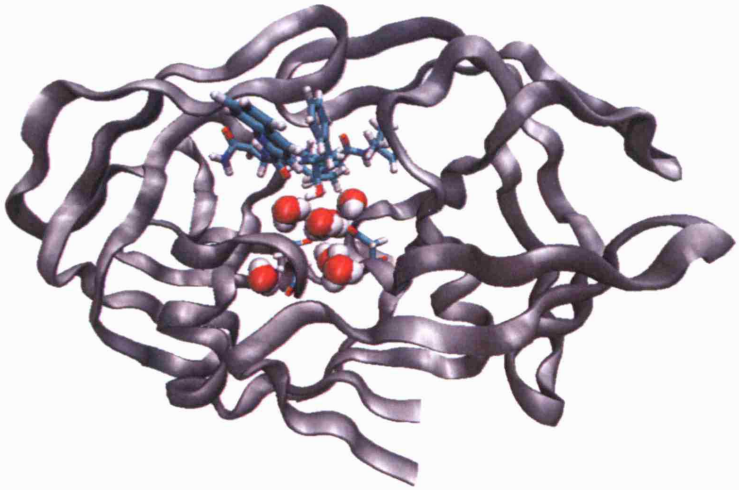
Molecular dynamics simulation studies of drug resistance in HIV-1 protease

Syed Kashif Sa'ad Ahmad Sadiq

<syed.sadiq@ucl.ac.uk>



A thesis submitted in fulfilment of the requirements
for the degree of Doctor of Philosophy

Department of Chemistry, University College London,
University of London,
2008

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

In the name of Allah, the Most Gracious, the Ever Merciful. [1]

Abstract

Overcoming the emergence of drug resistance in HIV is a major challenge to the scientific community. We use the established computational method of classical molecular dynamics to investigate the molecular basis of resistance in HIV-1 protease to the inhibitor saquinavir, using the wildtype and the G48V, L90M and G48V/L90M mutant HIV-1 proteases throughout this thesis.

Firstly we reveal insights into a G48V mutation-assisted lateral drug escape mechanism from the protease active site. Such a mechanism allows drug escape without the full opening of the flaps of the protease. Furthermore, the mechanism is facilitated by differential drug-protease interactions, induced by mutations that take advantage of the conformational flexibility of the inhibitor.

Secondly, we investigate the thermodynamic basis of binding of this set of mutants, using established 'approximate' free energy methods. The absolute and relative free energies of saquinavir binding to this set of proteases are successfully determined using our simulation and free energy analysis protocol and exhibit excellent correlation with experiment. This study is thus a template for an extended study on a larger range of HIV-1 protease-drug combinations. We describe a tool, the 'Binding Affinity Calculator', which has been designed to automate this protocol and which can be routinely applied, using high performance computing and grid technology, to meet the intensive computational demands of such an investigation.

The free energy of binding of the NC-p1 natural substrate cleaved by the protease is also determined. The enhanced flexibility of the substrate over the drug precludes the guarantee of a converged free energy result, even from the 10 ns duration of each simulation. However, qualitative insight into the thermodynamic basis of binding is gleaned as well as the effect of these mutations on the catalytic efficiency of the protease. Furthermore, we combine drug and substrate binding free energies to develop a metric for evaluating the approximate enzymatic fitness of a given mutant protease, computable directly from molecular simulation.

The *viva voce* examination was held on Wednesday 27 February 2008. The examiners were Dr. Charles Laughton and Professor Nora de Leeuw. One copy of this work has been deposited in the library of University College London. A second copy has been deposited in the University of London library at Senate House.

The picture on the title-page shows the effect of water ingress into the active site due to enhanced coupling to the flaps of mutant HIV-1 protease by the inhibitor saquinavir. The pictures in the bottom right of each page show the unsteered, followed by the steered, lateral expulsion of the inhibitor saquinavir out of the active site of the HIV-1 protease enzyme.

To my beloved parents,

Suraya and Matiullah.

Acknowledgements

الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ ۝٢

All praise belongs to Allah, Lord of all the worlds. [1]

Very little, if anything at all, is achievable completely by oneself. Were it not for the help and guidance of a great many people, the course that led to completing this difficult, but highly rewarding endeavour, would have been impossible to navigate.

I am forever grateful to my parents, Suraya and Matiullah, who have always lovingly guided, supported and encouraged me in all my pursuits.

I would also like to thank the late Hadhrat Mirza Tahir Ahmad for his inspirational guidance and wisdom.

I thank Peter Coveney for introducing me to the fascinating world of self-organising systems, giving me the opportunity to explore my interests and for advising me throughout my PhD course.

I thank my family, relatives and friends for their love, encouragement and prayers. I particularly thank my brother Tariq for teaching me to question, and who one day set a young boy on an irreversible journey by introducing relativity to him. Thank you Ellienurá for your loving and continuous support, without which this thesis would not have been completed. Thank you also Catherine and Iain for all your kind help.

I am grateful for the help of many colleagues, past and present, at the Centre for Computational Science. A special thanks to Shunzhou Wan, Ileana Stoica and Stefan Zasada with whom I have collaborated during the course of my PhD. I also thank Simon Clifford, Radhika Saksena, Daniel Scott, Shantenu Jha, Philip Fowler, Matt Harvey, Jon Chin, Gianni De Fabritiis, Rafael Delgado Buscalioni, Mary-Ann Thyveetil, James Suter, Catherine Gale, David Wright, Owain Kenway and Nilufer Betik.

I would like to thank our collaborators Paul Kellam, Simon Watson and Deenan Pillay at the Centre for Virology, UCL. Thanks also to Tanja van Mourik, Ben Moore, Robert Gifford, Richard Myers, David Wild, Annemie Vandamme, Koen Deforche and Charles Boucher.

I am indebted to the Engineering and Physical Sciences Research Council for paying my stipend and the EU funded ViroLab project for additional assistance.

Published work

This thesis is the product of my own work, unless otherwise stated. It is based in part on work described in the following refereed/to be refereed publications.

- I. Stoica, S. K. Sadiq, C. V. Gale and P. V. Coveney, Virtual physiological human research initiative; the future for rational HIV treatment design? *Submitted for publication.*
- S. K. Sadiq, M. D. Mazzeo, S. J. Zasada, S. Manos, I. Stoica, C. V. Gale, S. J. Watson, S. J. Zasada, P. Kellam S. Brew and P. V. Coveney. Patient-specific simulation as a basis for clinical decision making. *Submitted for publication.*
- S. K. Sadiq, S. J. Zasada, D. Wright, I. Stoica and P. V. Coveney. An automated molecular simulation-based binding affinity calculator for ligand-bound HIV-1 proteases. *Submitted for publication.*
- I. Stoica, S. K. Sadiq and P. V. Coveney (2008). Rapid and accurate prediction of binding free energies for saquinavir-bound HIV-1 proteases. *Journal of the American Chemical Society*, 130, 2639-2648.
- S. K. Sadiq, S. Wan and P. V. Coveney (2007). Insights into a mutation-assisted lateral drug escape mechanism from the HIV-1 protease active site. *Biochemistry*, 46, 14865-14877.
- S. K. Sadiq, S. J. Zasada and P. V. Coveney (2006). Grid assisted ensemble molecular dynamics simulations of HIV-1 proteases reveal novel conformations of the inhibitor saquinavir. *Lecture Notes in Computer Science*, LNBI 4216, Berthold, M.R., Glen, R. and Fischer I. (eds.) CompLife 2006, Springer-Verlag, pp.150–161.
- S. K. Sadiq, S. Wan, and P. V. Coveney (2006). Ensemble molecular dynamics of HIV-1 protease with the inhibitor saquinavir: Insights into the molecular basis of drug resistance caused by the G48V and L90M mutations. *Antiviral Therapy* 11:S151.
- P. V. Coveney, S. K. Sadiq, R. S. Saksena, M. Thyveetil, S. J. Zasada, M. McKeown and S. Pickles (2006). A lightweight application hosting environment for grid computing. *Proceedings of the UK e-Science All Hands Meeting.* pp.217–214. URL `http://www.allhands.org.uk/2006/proceedings/papers/675.pdf`.
- P. V. Coveney, S. K. Sadiq, R. S. Saksena, S. J. Zasada (2006). Constructing chained molecular dynamics simulations of HIV-1 protease using the application hosting environment. *Proceedings of the UK e-Science All Hands Meeting.* pp.428–431. URL `http://www.allhands.org.uk/2006/proceedings/papers/696.pdf`.

"Just as both tragedy and comedy can be written by using the same letters of the alphabet, the vast variety of events in this world can be realized by the same atoms through their different arrangements and movements."

– Werner Heisenberg [2]

Contents

List of Figures

List of Tables

Preface

BIOLOGICAL entities represent one of the most sophisticated and complex examples of self-organising systems, operating on a diverse set of spatiotemporal scales and evading the onset of entropy through the interactions that they mediate. Proteins are one of the fundamental components of all eukaryotic and prokaryotic biological systems. The unparallelled diversity of their potential structural make-up has allowed them to evolve to fulfil a vast array of functions within viruses, bacteria and multi-cellular organisms, all of which compete in an ever evolving ecosystem. Understanding protein function is the key to comprehending the complexity of interactions in biological systems. The functional complexity of proteins stem not only from the diversity of structures that they form but also from the subsequent variation in their dynamical properties. The 'holy grail' in molecular biology is therefore the ability to predict the structure and dynamics of proteins from their primary sequence of amino acids and thus infer their function.

In Chapter 1 we look at the structure, function and diversity of proteins and discuss the complexity of the protein structure prediction problem. We concentrate on the experimental methods that have been developed to discern protein structure, as well as looking at methods for evaluating the strength of the biochemical associations of proteins with ligands. In Chapter 2 we focus in detail on the computational method of molecular dynamics, based on the principles of classical physics and statistical mechanics, and its ability to study the structure and dynamics of biomolecular systems as well as its ability to furnish a route to the determination of the strengths of ligand-protein binding events.

From the interventional perspective of the medical sciences and the pharmaceutical industry, understanding protein dynamics, as well as structure, is the key to manipulating the biochemical processes of the body through the development of inhibitors of protein function that have the ability to impede pathogenicity within human beings. One of the most menacing examples of our time is the pathogenicity of the human immunodeficiency virus (HIV). Since 1981, around 25 million people have died due to AIDS, around 3 million in 2006 alone[1]. Overcoming this infectious virus, bolstered by its host of complex, evolving protein interactions that severely compromise the human immune system is a great and very real challenge to the scientific community. An account of the life-cycle of HIV and the development of AIDS along with the approach towards the medicinal treatment of HIV, is given in Chapter 3.

[1] http://www.unaids.org

We also focus on a discussion of the HIV protease enzyme, responsible for the maturation of the virus, looking specifically at its molecular structure and function and the problem of emergent drug resistant mutants of the protease that limit the efficacy of anti-retroviral inhibitors (ARVs) designed to bind to it.

In Chapters 4 and 5, we investigate by means of molecular dynamics (MD) simulations, the kinetic basis of drug resistance conferred by a set of drug resistant mutants of HIV-1 protease bound to a specific inhibitor. Whilst the study reported in Chapter 4 utilises MD methods to explore the conformational flexibility of the inhibitor in the active site of the protease, leading to differential binding, in Chapter 5, we report the longer timescale effects of differential drug binding in the active site, which leads to the identification of a mutation-assisted drug expulsion mechanism.

In Chapter 6, we investigate the thermodynamic basis of drug resistance for the same set of mutants and the same inhibitor, again using MD simulations. We report a protocol for the accurate determination of the absolute binding free energy between protease and inhibitor as well as the correct ranking of the binding strengths of drug resistant mutants, using 'approximate' free energy methods. Furthermore, the use of MD permits a molecular level insight into the cause of the observed differences in thermodynamic binding affinity. As the construction and execution of MD simulations is an involved and laborious task, we have additionally developed a tool, the Binding Affinity Calculator (BAC), for the automation of simulations as well as binding free energy calculations of HIV-1 protease-ligand variants. Mentioned briefly in Chapter 6, the BAC is described more fully in Appendix A.

In Chapter 7, we focus our attention not on an inhibitor of the protease but on one of its naturally processed substrates. We again employ MD simulations and the free energy methods reported in Chapter 6, facilitated by the use of the BAC, to study the strength of binding of mutant HIV-1 proteases with this natural substrate. We are thus able to provide molecular insight into the effect of drug resistant mutations on the catalytic efficiency of the protease. Moreover, we discuss how a suitable combination of the changes in both drug resistance and catalytic efficiency provide a better description of the overall enzymatic fitness of a particular mutant strain of the protease in the presence of an inhibitor.

We conclude in Chapter 8 with a view towards future research that may prove useful in the further study of HIV-1 protease and HIV, in general. A complete bibliography is provided at the end of the thesis.

CHAPTER 1

Experimental Methods for Investigating Proteins

PROTEINS are the most diverse class of macromolecules in any biological system. They are organic heteropolymers made from combinations of the 20 standard amino acids that exist in nature. Their role in the body spans all levels of intracellular and extracellular processes. Proteins come in all range of sizes, although many are between 200 - 300 amino acids in length. Some are very small (only a few amino acids long) and are called peptides, whilst the largest known protein, 'titin', found in skeletal and cardiac muscle contains 26,926 amino acids in a single chain. Proteins mediate important chemical reactions as enzymes, as well as being transporters for essential molecules. They form the pores in membranes, selectively allowing molecules into and out of cells and they are essential to the complex signalling pathways the body utilises for self-regulation, to mention but a few of their vast array of functions.

One of the principal mechanisms by which many proteins function is through association or binding with other proteins and/or ligands. Determining the strength of such binding events is important for a quantitative understanding of biological interactions. Enzymes in particular, tend to have biochemically active regions called 'active sites' into which, usually smaller, organic structures such as peptides, tend to fit. Proteins are of special interest to the pharmaceutical industry. Interfering with the normal processing activity of proteins such as enzymes, through the development of molecular inhibitors designed to selectively bind to targets, is vital for the successful and safe application of drugs in the medical environment.

In this chapter we will explore the nature of proteins, their structure and function, the importance and advantage of understanding their behaviour and the complexity of determining their structures and functions. Of the various methods adopted for the determination of protein structure, two will be described in this chapter. These are X-ray crystallography and Nuclear Magnetic Resonance (NMR). We will also look at some of the experimental methods employed to evaluate the strength of proteins binding with other molecular structures such as inhibitors and substrates as well as outlining the principles of enzyme kinetics and inhibition in relation to the strengths of binding of ligands. A thorough treatment of this subject can be found in several standard texts on molecular biology [3–6].

1.1 Protein Structure and Formation

Proteins consist of long chains of a set of repeating units called amino acids. There are 20 standard amino acids found in nature (see Fig 1.1), all with differing molecular shapes and structures. However they all share a common backbone with a central carbon atom (C_α) bonded to an amine group (NH_2), a carboxylic acid group (COOH), a hydrogen atom and a group R which defines the amino acid type and is either hydrophobic, hydrophilic or charged. Amino acids are also stereo-isomers except for glycine and can therefore have right handed or left handed forms. It is interesting that biological systems predominantly use the L-form (see Figure 1.2).

Amino acids combine in linear chains by condensing to form peptide bonds between their amine groups and the carboxylic groups of successive amino acids. This forms a repeated backbone structure of N, C_α and C atoms with a side-chain (R) attached to each C_α atom. The ends of the amino acid chain are called the N-terminal and C-terminal and are usually not condensed.

This linear formation of hetero-polymeric chains is regarded as the primary structure of a protein. A protein then adopts secondary structure, characterised by the formation of distinct coils in the chain known as α-helices or of sheets of turned parallelised chains known as β-sheets. Both these structures are stabilised by the formation of hydrogen bonds between amino acid residues (see Fig 1.3).

Following this, the polypeptide chain begins to 'fold' into a three dimensional structure which is characterised by the packing of its α-helices and β-sheets into a myriad of different globular shapes. This is now the tertiary structure of a protein and is achieved at slower timescales (μs to ms) than the formation of secondary structure. More complex proteins are not limited to one polypeptide chain but can be formed from the folding of several chains into compact domains which further associate non-covalently to form more complex structures. These are termed quaternary structures. Through this complex process of folding, residues that were initially far away from each other in sequence are brought within spatial proximity to each other [5].

The primary sequence of a protein is encoded by DNA. DNA is also a heteropolymer, but made of only four types of repeating units called bases (A, T, C and G). In cells, DNA is transcribed into RNA by a protein called RNA-polymerase and this is then translated into an amino acid sequence by ribosomes. Three consecutive bases on an RNA chain form a codon, which is responsible for encoding one amino acid. As there are 64 possible codons and only 20 amino acids there is a lot of redundancy which has allowed some codons to signal the beginning or end of polypeptide translation (see Figure 1.4).

Once the polypeptide chain is formed, it then folds either unaided [7] or through the help of yet more proteins called chaperones. It is possible for errors to occur in the protein folding process. A protein may reach an intermediate state and then progress in an incorrect folding direction leading to 'misfolding'. This can lead to the phenomenon of amyloidosis, characterised by the formation of misfolded aggregated insoluble proteins and is responsible for diseases such as Alzheimer's and Creutzfeld Jacob Disease

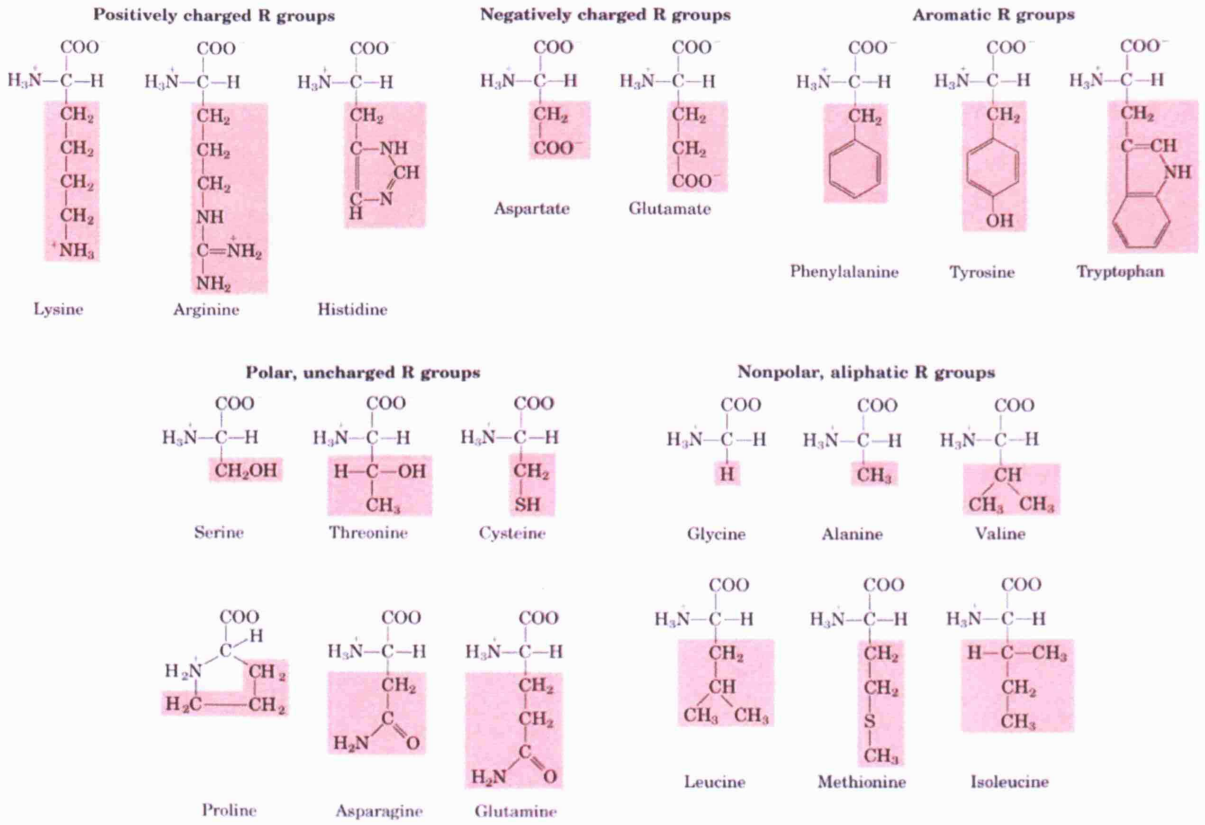


Figure 1.1: The twenty standard amino acids that exist in nature (reproduced from: http://www.le.ac.uk/by/teach/biochemweb/tutorials/aminoacidstruct.html).

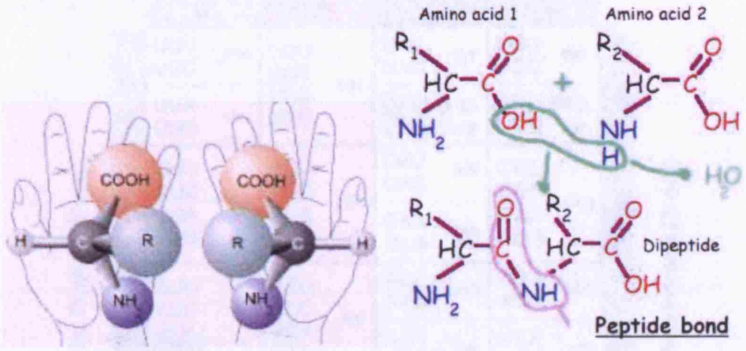


Figure 1.2: Left - The stereo isomers of amino acids. Nature predominantly uses the L-form. Right - Amino acids link into chains by forming peptide bonds and release water molecules in the process (reproduced from: http://www.genomeprairie.ca/enablingtech/spectrum/toolbox/boyd2.htm).

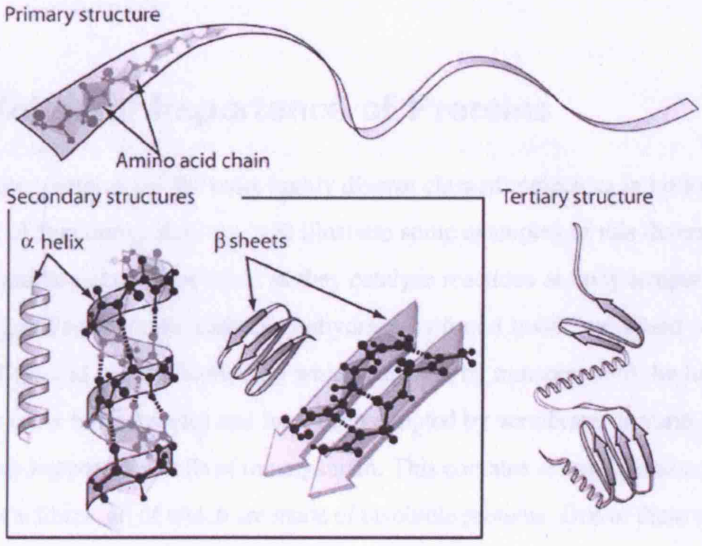
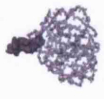


Figure 1.3: The primary, secondary and tertiary structures of proteins. α-helices are formed by amino acids able to form hydrogen bonds within the coils. β-sheets are formed by hydrogen bonding between the strands (reproduced from: http://www.press.uillinois.edu/epub/books/brown/ch6.html).



(CJD).

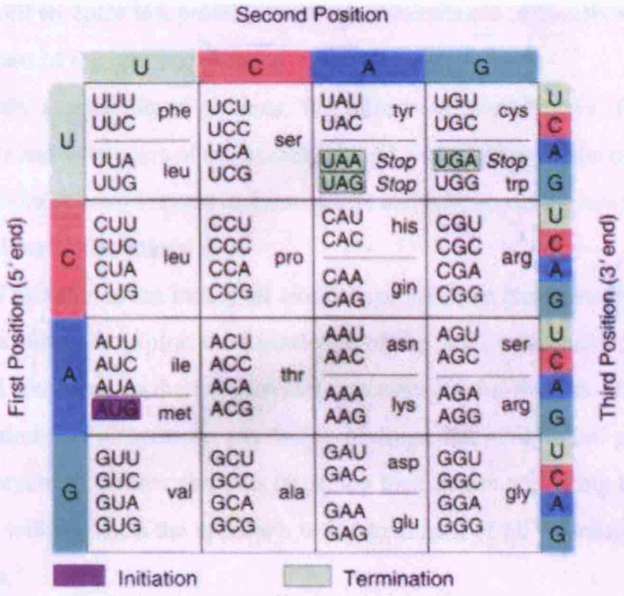
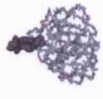


Figure 1.4: The coding of amino acids in RNA. Three of the 4 bases (U, A, C, G) make up one codon. There are therefore 64 codons that code for 20 amino acids and thus a lot of redundancy (reproduced from: http://chuma.cas.usf.edu/ karl/evolution/chapter_4.htm).

1.2 The Role and Importance of Proteins

As mentioned before, proteins are the most highly diverse class of molecules in biological systems and have a wide range of functions. Here we will illustrate some examples of this diversity. Enzymes are one of the most important class of proteins as they catalyse reactions at body temperature by lowering the activation energy. For example, carbonic anhydrase, is found inside red blood cells and catalyses the conversion of CO_2 and water into H_2CO_3 which can then be transported to the lungs. Proteins can also associate with other biomolecules and have been adopted by vertebrates to form the extra-cellular-matrix (ECM) which supports the cells of the organism. This contains connective tissue such as tendons, ligaments and muscle fibres, all of which are made of insoluble proteins. One of these proteins, collagen, is the most abundant protein in the human body.

The haemoglobin molecule in red blood cells is an example of a protein used to carry crucial molecules around the body. Each heme group containing one iron atom at its centre, binds to one oxygen molecule and thus facilitates the transport of oxygen around the body. Some hormones are also proteins as well as hormone receptors, and thus proteins are intrinsic in the signalling and communication of cells with each other. Proteins can also exist on and through cell membranes acting as selective



gateways into and out of the cell. For example aquaporins control the flow of water into and out of the cell [8], whilst the T-cell receptor is a protein on the outer membrane of T-cells that recognises certain peptide fragments as part of the immune defence.

Viruses also contain a selection of proteins, for various enzymatic tasks, forming the structural matrix of viral particles and as triggers of membrane fusion by which they can be incorporated into target cells. We will elaborate on this with regard to the structure and function of the human immunodeficiency virus in greater detail later in this thesis.

The importance of proteins to the industrial world, especially the pharmaceutical industry, is therefore self evident. The ability to exploit the knowledge of the functional characteristics of proteins to design molecules that intervene in the biochemical processes of the body is of prime importance to the pharmaceutical industry. Furthermore, the design of drugs that inhibit viral proteins as a means of intervening in the lifecycle of viruses and thus impeding their action are taking hold progressively. In Chapter 3, this thesis will focus on the approach to the treatment of HIV through the development of viral protein inhibitors.

1.3 Sequence-Structure-Function Relationships

The functional properties of proteins fundamentally depends on their three-dimensional structure. It is the myriad of shapes that different proteins fold into that provides them with the diversity and specificity to carry out their many functions. The vast number of possible sequences that a protein can be composed of, once folded into a unique three dimensional structure, allows for this incredible specificity of function. The ideal goal then in molecular biology for determining the function of proteins would be the ability to deduce or predict the three dimensional structure of proteins from their amino acid sequence. The inability of the scientific community to solve this has come to be known as the 'protein folding problem'.

Many attempts have been made to overcome the complexity of the protein folding problem. Essentially an amino acid chain folds due to the dynamical laws of physics between its constituent parts. The ability to model this computationally is currently beyond our scope. Experimental methods such as X-ray crystallography and NMR spectroscopy have provided a means to determine the structures of certain proteins but these processes are rather slow. Unfortunately the number of solved structures is rather limited (around 40000). In contrast high throughput methods have emerged that have vastly increased the rate at which genomes are being sequenced. There is therefore still a divergence between the rates at which sequences and structures are determined. Bioinformatical methods have attempted to discern unknown protein structures of known sequences by matching with sequences of known structure. This is known as homology modelling and has been met with limited success as its accuracy is governed by the numbers of available matching structures [9].

Proteins generally fold into the structure that corresponds to their global minimum of free energy under physiological conditions. What is interesting is that the free energy difference between unfolded and folded structures is very small, usually between 5-15 kcal/mol. This is due to the decreased enthalpic and decreased entropic changes from an unfolded to a folded state compensating each other (Equation 1.2). The mechanism by which a protein folds is of considerable interest and was considered by Levinthal in 1968, who showed that a protein could not fold into its equilibrium structure through an exhaustive search of all of its possible conformations. Consider a protein with 100 amino acids. If each amino acid has two possible conformations then the number of possible conformations of the protein is 2^{100}. If interconversion between conformations took just 10^{-11} seconds, then it would still take longer than the estimated age of the universe for this simple protein to fold. As proteins fold generally on the μs to ms level, there is a discrepancy and this has come to be known as the *Levinthal Paradox* [10]. It is therefore evident that proteins do not fold by randomly sampling every conformation available to them and instead must follow specific pathways, being guided by the physical interactions of their constituent amino acids as well as their environment.

The dynamical properties of proteins also determine their function. It is essential to regard proteins as dynamic entities that are never static even at equilibrium. Comprehending this dynamics will aid not only the understanding of protein folding, which is a far from equilibrium process, but in the fluctuations of proteins once they have arrived at their equilibrium folded structures. Often it is the subtle differences in equilibrium dynamics that differentiate proteins that are almost structurally identical. A great example of this is the emergence of drug resistant strains of particular proteins through sometimes single mutational changes in the amino acid sequence that results in almost no structural difference once folded. We will look in great detail into this phenomenon later in the thesis with regard to the emergent drug resistance of HIV viral proteins (see Chapter 3).

1.4 Experimental Methods for Structure Prediction

There are two main techniques used to determine protein structure, X-ray crystallography and NMR spectroscopy. X-ray crystallography has historically been the older and the most popular of the two. The first protein structure to be determined was that of myoglobin in 1958 using X-ray crystallography. The consolidation of all determined structures has been achieved by means of the Protein Data Bank[1], a web-based database containing all currently resolved protein structures [11]. By June 2007, 43873 structures had been deposited in the Protein Data Bank of which 37316 were from X-ray crystallography and 6320 from NMR. A thorough treatment of these techniques can be found in the literature [5].

[1] `http://www.rcsb.org/pdb`

1.4.1 X-ray Crystallography

X-ray crystallography works through the analysis of the diffraction patterns formed when protein crystals are illuminated with X-ray beams. When beams from an X-ray source strike a crystal, most travel straight through, but some are scattered by the electrons in the atoms of the crystal. Most of these scattered rays cancel, but some interfere constructively and form diffraction patterns that can be recorded on a photographic plate. Interference is essentially due to the path difference set up in the beams that scatter off different atoms in the crystal lattice. Bragg's law of diffraction can then be employed to relate the spacings in the diffraction patterns with the atomic spacings in the lattice and thus determine the crystal structure. The amplitudes and the phases of the diffraction data are then used to produce an electron density map of the repeating unit of the crystal.

The problem with employing this method is the requirement of well-ordered crystals to have formed prior to the application of X-rays. Unfortunately, it is very difficult to obtain ordered protein crystals and therefore difficult to obtain diffraction patterns that can be sensibly interpreted. Other problems are that crystal growth can be slow and the formation of crystals is critically dependent on a range of different parameters such as pH, temperature, protein concentration, ionic or ligand presence and several other factors. The effect of these factors means that protein crystallisation can sometimes take months to achieve and may also sometimes not be possible.

The more ordered the crystal, the higher the resolution of diffraction data and in turn, the higher the quality of the electron density map and the smaller the error in the protein structure prediction. At 5 Å the shape of the molecule can be obtained; at around 3 Å, the path of the polypeptide chain can be traced and at 1 Å the atoms can be distinguished as discrete balls of density [5].

The temperature factor (B-factor) is also discerned from the X-ray data and gives a measure of the positional error, but also may indicate the flexibility of that part of the structure. Unfortunately, it cannot distinguish between these two possibilities.

1.4.2 Nuclear Magnetic Resonance (NMR)

NMR exploits the quantum property of spin (magnetic moment) possessed by certain atomic nuclei such as ^{1}H, ^{13}C, ^{15}N and ^{31}P to probe their exact chemical environment and can be used to determine the distances between these atoms and others in a molecule. These distances are subsequently used to construct a three dimensional model of protein structure.

NMR is facilitated by the abundance of the carbon, hydrogen and nitrogen atoms in proteins and for this reason ^{1}H atoms are usually analysed to determine the structure of small proteins. Larger proteins can make use of ^{13}C and ^{15}N which can be incorporated into proteins grown in media that are isotopically rich.

When ^{1}H atoms are placed in a strong magnetic field their spin aligns along the field as they accom-

modate low energy configurations. Applying radio frequency (RF) pulses to the atoms allows them to absorb photons and promotes them into higher energy states. The frequency at which this occurs is the resonance frequency ν of that particular atom in its chemical environment and depends on the strength of the effective magnetic field of the nucleus (B), see (Equation 1.1):

$$\nu = \gamma \mathbf{B} \tag{1.1}$$

where $\gamma = -e/2m$ is the gyromagnetic ratio of the particle. The effective magnetic field of the particle is sensitive to the exact chemical environment and it is this that allows ^{1}H atoms to be differentiated from each other. The resonance frequency is measured relative to a standard molecule and the shift in frequency away from this is called the *chemical shift* and measured in parts per million (ppm).

In practice, the conformation of a molecule is determined by an NMR method known as 2D-nuclear Overhauser enhancement spectroscopy (2D-NOESY). NOESY allows the distances between atoms close to each other to be determined from the chemical shift and so is effective in tertiary structure determination. Other methods such as 2D-correlated spectroscopy (2D-COSY) provide information on atoms which are covalently separated by a number of bonds. This allows secondary structure information to be determined.

The difficulties associated with using NMR for protein structure determination include the requirement of high concentrations of proteins (1-2 mM) at a maximum of pH 6. This makes many proteins susceptible to aggregation and is also not a physiologically representative situation. Further to this, only small proteins have yet been determined, the upper limit currently being proteins that weigh 25 kDa, where 16 Daltons (Da) is the weight of 1 oxygen atom.

1.5 Enzyme Kinetics and Inhibition

Enzymes are proteins that catalyse biochemical reactions. They are essential in effecting and regulating the complex and diverse networks of reactions that occur in biological organisms and, as such, their correct functioning is important in the maintenance of their host organism.

There are several classes of enzyme, depending on the types of reaction they catalyse. Amongst these are oxidoreductases, which catalyse redox reactions, transferases, which catalyse reactions in which one chemical group is transferred between substrates, and hydrolases, which catalyse hydrolysis reactions, as well as several others [6].

Enzymes enhance the rates of chemical reactions between $10^6 - 10^{14}$ times over the uncatalysed reaction and furthermore are very selective about the substrates with which they interact. Enzymatic rate enhancement is variable, being dependent on the enzymes' physical and chemical environment and can vary greatly with alteration in temperature and pH. Such selectivity ensures that adverse reactions

are not catalysed and this, coupled with the environment-dependent rate enhancement, is the basis of how all enzymes work.

Due to their crucial role in sustaining biological life, enzymes are of great importance to the pharmaceutical industry and often serve as the targets for inhibitors, designed to interfere with their working. One of the main means of relieving pain and inflammation is through the application of a class of inhibitors known as non-steroidal anti-inflammatory drugs (NSAIDs). For example, prostaglandin H2 synthase, which synthesises a prostaglandin precursor known as PGH_2, is involved in local pain response and inflammation and is inhibited by several NSAIDs such as ibruprofen and aspirin [12, 13].

Inhibitor design is by no means limited to human enzymes or proteins. The proteins of infectious organisms such as bacteria and viruses are also important targets for the pharmaceutical industry. One important example is HIV, for which several anti-retroviral inhibitors (ARVs) have been developed to impede the function of a range of viral enzymes. The inhibition of enzymes of HIV is discussed in detail in Chapter 3.

1.5.1 Thermodynamic Considerations and the Concept of Binding Affinity

The strength with which proteins bind to each other, with substrates, inhibitors or other ligands is measurable by a quantity known as the 'binding affinity' of such association events. Due to the extensive association of biochemical species in nature, the binding affinity of two biochemical species is arguably one of the most important metrics to ascertain in such systems.

Binding affinities are expressed differently according to the variety of methods implemented to measure them. The fundamental description of binding affinity stems from the thermodynamic principles of free energy. For a full treatment of free energy in equilibrium thermodynamics, the reader is referred to the literature [14]. The concept of free energy will also be revisited in Chapter 2.

The thermodynamic free energy of a system is a measure of the available work that can be extracted from it. There are several types of free energy, each corresponding to the set of thermodynamic variables that remain constant as the system interacts.

When two reactants interact at constant temperature and pressure, there is a change in the free energy of the entire system, given by Equation 1.2:

$$\Delta G = \Delta H - T\Delta S \tag{1.2}$$

ΔG is the Gibbs free energy change of the system at temperature T, and is composed of changes in the enthalpy (ΔH) and entropy (ΔS) .

Understanding the difference in free energy between reactants and products allows us to determine whether a reaction will occur spontaneously or not. Only if the free energy of the pure products is smaller than that of the pure reactants, can a reaction occur spontaneously, and will do so until equilibrium is

reached, whereby the free energy of the system is minimised. If the reactants bind together, the reaction can be written as:

$$A + B \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} AB \tag{1.3}$$

where A and B bind to form the complex AB and k_1 and k_{-1} are the rate constants of the forward and backward reactions respectively.

At equilibrium, the rate of the forward and backward reactions are equal such that there is a mixture of species A, B and the complex AB. The concentrations of each of these species at equilibrium then determine the equilibrium binding constant, K and the equilibrium dissociation constant, K_d, given by:

$$K = \frac{1}{K_d} = \frac{k_1}{k_{-1}} = \frac{[AB]}{[A][B]} \tag{1.4}$$

where $[A]$, $[B]$ and $[AB]$ are the concentrations of each species. K_d is thus a measure of the 'binding affinity' of such a reaction, the smaller the value of K_d, the greater the binding affinity and the greater the concentration of the complex AB at equilibrium.

An alternative measure of the binding affinity is the free energy difference of binding ΔG_b, given by:

$$\Delta G_b = G(AB) - G(A) - G(B) \tag{1.5}$$

where the free energies of pure product $G(AB)$ and reactants $G(A)$ and $G(B)$ are usually measured in kcal/mol. The greater the negative value of ΔG_b, the stronger the binding affinity between species A and B.

However, at equilibrium there is a direct relationship between the free energy difference of binding ΔG_b and the dissociation constant K_d, given by the *van't Hoff* equation (Equation 1.6):

$$\Delta G_b = -RT\,lnK \tag{1.6}$$

where R is the gas constant and T is the temperature. Equation 1.6 is thus very important. It states that the position of equilibrium in a binding reaction, at a given temperature, is dependent only on the difference in free energy of binding between the pure product and the pure reactant and not the difference in free energy of the system between the beginning and the point at which equilibrium is reached. Therefore, the more negative the value of ΔG_b, the more a reaction will side towards a greater concentration of complex.

1.5.2 Kinetic Basis of Enzyme Catalysis and Inhibition

Whilst a thermodynamic approach alone allows an understanding of the eventual position of equilibrium, it yields no information about how quickly a reaction proceeds. The rate of a reaction is governed

by its 'activation energy'. For any given temperature, this is the free energy difference between the reactants and the transition state. In principle, therefore, even though a reaction may be thermodynamically favourable, a very high activation barrier would result in an extremely slow approach to equilibrium.

Enzyme Catalysis

The physicochemical basis of enzymatic and indeed catalytic function generally lies in the reduction of the activation energy required for a reaction to proceed. The only available energy an enzyme can make use of to lower this barrier comes from the binding affinity between the reactant substrates and the enzyme. The binding affinity corresponds to the change in free energy of the enzyme-substrate complex upon binding, the larger this free energy difference is, the stronger the binding affinity is and the more energy is available for the enzyme to lower the activation energy of the reaction. To demonstrate this, let us consider the kinetics of a single-substrate enzyme reaction, where the enzyme (E) binds to a substrate (S) to form a product (P):

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P \tag{1.7}$$

where k_1 and k_{-1} are the rate constants of substrate binding and dissociating respectively from the enzyme and k_2 is the rate constant of the product forming step of the reaction. The rate of the reaction (v) is given by the *Michaelis-Menten* equation:

$$v = \frac{V_{max}[S]}{K_m + [S]} \tag{1.8}$$

where K_m is the *Michaelis constant* for the reaction, $[S]$ is the free substrate concentration and V_{max} is the theoretical maximum rate of the reaction, which occurs if all the enzyme molecules are saturated by the substrate such that the concentration of bound enzymes $[ES]$ equals the total enzymatic concentration $[E]_{tot}$:

$$V_{max} = k_2[E]_{tot} \tag{1.9}$$

There are essentially two ways of deriving the above rate equation, depending on the assumptions that are made about the reaction and these give rise to differences in the meaning of K_m. In the 'steady-state' approximation, the concentration of the enzyme-substrate complex $[ES]$ is assumed to be constant. Furthermore, it is assumed that the complex intermediate is short-lived such that $k_2 >> k_1$. In the steady-state approximation, K_m is given by:

$$K_m = \frac{k_{-1} + k_2}{k_1} \tag{1.10}$$

In the 'equilibrium' approximation, it is assumed that the complex ES is in equilibrium with E and S. This is the same as assuming the steady-state approximation as well as the condition that $k_{-1} >> k_2$. In such a scheme, K_m becomes the dissociation constant of the enzyme-substrate complex $K_d^{E:S/ES}$:

$$K_m = K_d^{E:S/ES} = \frac{k_{-1}}{k_1} \tag{1.11}$$

An understanding of K_m is therefore important in ascertaining the rate at which the enzymatic reaction proceeds. In the equilibrium approximation, the stronger the binding affinity between enzyme and substrate, the faster the rate of the reaction will be.

Another important quantity to understand in enzymatic reactions is the *catalytic constant*, k_{cat}, which has units $M^{-1}s^{-1}$. This is the rate constant for the decomposition of the enzyme-substrate complex into the products and so, in the kinetic scheme used in Equation 1.7, it is the same as k_2. Together, the values of k_{cat} and K_m provide a measure of the overall rate of a particular enzymatic reaction as well as the selectivity of one enzyme compared to another. As such they are determined extensively in experimental enzymatic studies, where often the *specificity constant*, $k_A = k_{cat}/K_m$ is quoted. The larger the value of k_A, the faster the rate of the reaction for any given substrate concentration $[S]$.

Enzyme Inhibition

When an inhibitor is introduced into a mixture of enzyme and substrate, the kinetic scheme alters depending on the method of inhibitor interference. The main methods of enzymatic inhibition are competitive, non-competitive and mixed. A competitive inhibitor usually binds to the same enzymatic site as a substrate and thus increases K_m, leaving k_{cat} unaltered. A non-competitive inhibitor binds to a completely different site and thus reduces k_{cat}, leaving K_m unaltered. A mixed inhibitor reduces k_{cat} and may increase or decrease K_m due to a partial overlap with the binding site of the substrate as well as sometimes being part of the binding site. Here we will discuss the kinetic scheme for competitive inhibition where it is assumed that the enzyme cannot bind the substrate and inhibitor simultaneously:

$$E + S + I \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES + I \xrightarrow{k_2} E + P + I \tag{1.12}$$

$$E + S + I \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} EI + S \tag{1.13}$$

The rate constants k_1, k_{-1} and k_2 are represented the same as in Equation 1.7, whilst there is an additional reversible reaction for the binding of the inhibitor I to the enzyme, for which the forward and backward rate constants are k_3 and k_{-3} respectively. Both the steady state and equilibrium approximations yield the same rate equation for such a reaction scheme, which is a modified Michaelis-Menten equation:

$$v = \frac{V_{max}[S]}{K_m\left(1 + \frac{[I]}{K_i}\right) + [S]} \tag{1.14}$$

where K_m again depends on the approximation used (see Equations 1.10 and 1.11, $[I]$ is the inhibitor concentration and K_i is the equilibrium *inhibition constant* of the enzyme-inhibitor complex $K_d^{E:I/EI}$:

$$K_i = K_d^{E:I/EI} = \frac{k_{-3}}{k_3} \tag{1.15}$$

K_i is therefore an important metric in understanding enzyme inhibition and has units M. The smaller the value of K_i the stronger the binding affinity between enzyme and inhibitor and the greater the decrease in the enzymatic reaction rate. As K_i is directly linked to the free energy difference of binding of the inhibitor to the enzyme, via Equation 1.6, maximising ΔG_{bind} is the goal of inhibitors designed to impede enzymatic function.

1.5.3 Experimental Methods for Determining Binding Affinities

Experimentally it is impossible to measure the free energy change at the molecular level of protein-ligand binding. In general therefore, experiments are devised that follow the reaction kinetics of mixing ligand to protein. We will describe two techniques by which this can be done, namely enzyme assays and isothermal titration calorimetry, concentrating on a description of enzyme activity and inhibition which is related to the rest of this thesis.

Enzyme Activity and Inhibition Assays

Conducting enzyme inhibition assays (EIA) is a key method by which parameters such as K_m, k_{cat} and K_i can be determined. In general, enzyme assays monitor the production of product or the consumption of substrate used in a reaction. There are also several types of assaying experiment, varying according to the stage of an enzymatic reaction being probed. We will discuss the steady-state and pre-steady-state experimental methods here.

When a substrate is added to an enzyme, there is an initial transient stage, known as the pre-steady-state, in which the concentration of the enzyme-substrate complex increases sharply due to rapid uptake of the substrate by the free enzyme. This is followed by the steady-state regime, during which the complex concentration is approximately constant for a limited period of time. The steady-state kinetic scheme has been described in § 1.5.2. Experimentally, the reaction kinetics in the steady-state scheme can be followed by a range of methods. A common method is that of fluorometric assaying in which the difference between the fluorescence of the substrate and the product is used as a metric for following the rate of production of the product. The values of K_m and k_{cat} can then be discerned by fitting the observed rate to the Michaelis-Menten equation (Equation 1.8).

Similarly, the strength of enzyme-inhibitor binding can then be determined by introducing the inhibitor and measuring the change in the enzymatic reaction rate upon inhibitor binding in the steady-state kinetic scheme. In this case the value of K_i can be determined by fitting the observed rate to the mod-

ified Michaelis-Menten equation (Equation 1.14). A common way of doing this is by measuring IC_{50} or IC_{90} values. These are the concentrations of inhibitor required to bind 50% or 90% of the enzyme respectively and thus the concentration required to impede the rate of reaction by the same percentage. If K_m is known, K_i can then be inferred from a recasting of the modified Michaelis-Menten equation in terms of the IC_{50} concentration:

$$K_i = \frac{IC_{50}}{\left(1 + \frac{[S]}{K_m}\right)} \tag{1.16}$$

However, often IC_{50}s are provided as an indirect measure of inhibitor strength. Whilst this is fine for comparing different inhibitors with the same enzyme, it is not strictly valid for comparing different enzymes with the same inhibitor. This is due to the possibility that different enzymes or mutant forms of the same enzyme will possess different enzymatic activities and will thus vary in K_m. Therefore, whether changes in the overall rate of a reaction in the presence of an inhibitor across a range of enzymes is due to a change in the drug-binding affinity or a change in the catalytic activity of the enzyme is not explicitly possible to distinguish using IC_{50} alone.

An alternative way of determining the strength of enzyme-inhibitor binding is by following the reaction kinetics of mixing only the enzyme and the inhibitor via pre-steady-state experiments. These are often harder to conduct than steady-state experiments as they require rapid mixing and observation techniques. The advantage of conducting pre-steady-state experiments however, is that they provide the forward and backward rate constants of the enzyme inhibition directly and not just the equilibrium inhibition constant and thus allow the kinetic mechanisms for a whole variety of enzyme inhibition reactions to be studied. The equilibrium between an enzyme and an inhibitor can be expressed by Equation 1.17:

$$E + I \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} EI \tag{1.17}$$

$$K_d = \frac{k_{off}}{k_{on}} \tag{1.18}$$

where k_{on} and k_{off} are the rate constants of the forward and reverse reactions respectively and K_d is the equilibrium inhibitor dissociation constant. The observed rate constant k_{obs} is then defined by:

$$k_{obs} = k_{on}[I] + k_{off} \tag{1.19}$$

where [I] is the concentration of inhibitor and is in excess compared to the enzyme concentration.

Pre-steady-state enzyme inhibition assays use protein fluorescence decay measurements to determine k_{obs} and thus follow the reaction kinetics. This only works for inhibitors that quench the protein fluorescence upon binding. Varying the concentration of inhibitor produces linear correlations with k_{obs} from which k_{on} can be determined. k_{off} can be measured by reacting the enzyme-inhibitor mixture with

another inhibitor that doesn't quench the fluorescence and this in turn allows K_d to be calculated, where K_i is approximated to K_d.

An excellent application of such techniques to HIV-1 protease can be found in the literature [15]. Finally, the *van't Hoff* equation (Equation 1.6) can be used to relate the free energy change of binding to K_i, although this last conversion is not generally made by experimentalists as binding affinity is usually compared in terms of the value of the dissociation constant K_i.

Isothermal Titration Calorimetry

Isothermal Titration Calorimetry (ITC) directly measures the energy associated with a chemical reaction, triggered by the mixing of two components. It is the only method currently available that can measure not only the binding affinity but also the enthalpic ΔH and entropic ΔS contributions to the free energy (see Equation 1.2).

ITC is conducted by the stepwise addition of one reactant to another in a reaction cell and measures the power necessary to maintain a constant temperature difference between the reaction and reference cells. Each injection releases or absorbs heat (q_i) and this is proportional to the amount of ligand that binds to the protein and the enthalpy of the reaction ΔH by the relation (Equation 1.20):

$$q_i = v\Delta H \Delta L_i \tag{1.20}$$

where v is the volume of the cell and ΔL_i is the increase in the concentration of bound ligand after the i^{th} injection. The energy required to maintain the same temperature decreases with each subsequent injection as less free reactant remains to become bound (see Figure 1.5). Data analysis yields both ΔH and the binding constant K_d which in turn gives ΔG (Equation 1.6). Through the standard thermodynamic relationship (Equation 1.2), ΔS can then be calculated.

The advantage of this method is its ability not only to compare binding affinities of different enzyme-inhibitor complexes but also to calculate the enthalpic and entropic contributions too. ITC allows us to distinguish between two complexes whose binding affinity is the same even though they are achieved through a different balance of enthalpic and entropic contributions. This can in turn confer added information on the specific advantage or disadvantage of certain inhibitors.

In recent years, the binding properties of many ligands to targets have been deposited in a web-accessible database, known as 'BindingDB'[2] [16, 17]. The database contains information on EIA, ITC and IC_{50}s as well as supplementary information and links to the Protein Data Bank and other resources. In Chapter 3 we will discuss the use of the techniques described here in the context of the enzymatic and inhibition properties of ligand-bound HIV-1 proteases.

[2] `http://www.bindingdb.org`

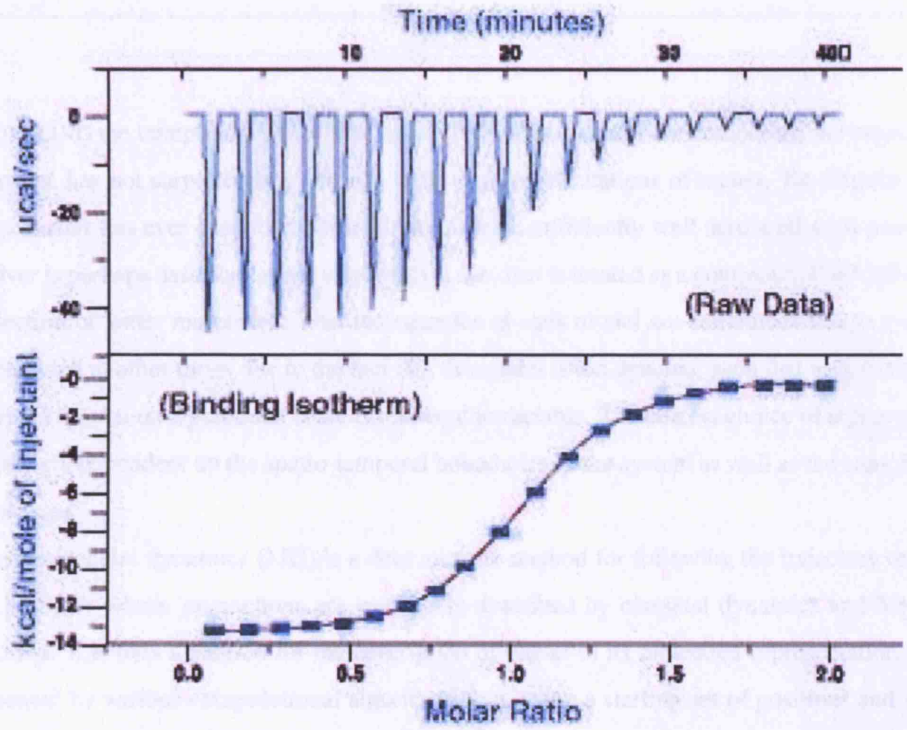


Figure 1.5: A typical graph from an ITC measurement. As more of the reactants are mixed in discrete injections, the compensatory power spikes required to maintain a constant temperature become less pronounced. The area under the peaks is then the heat produced by the reaction and can be used to calculate the enthalpy (Δ H) and K_d giving both the free energy change and thus the entropic (ΔS) contribution too (adapted from: http://biophysics.uchicago.edu/calorimetry.htm).

CHAPTER 2

The Molecular Dynamics Toolbox

MODELLING the complex physical world, which extends spatially and temporally across so many scales, has not surprisingly given rise to several representations of matter. No singular representation of matter has ever described physical interactions sufficiently well across all such scales; the flow of a river is perhaps described more effectively if the river is treated as a continuum fluid rather than a vast collection of water molecules. The inadequacies of each model are sometimes due to a lack of specific detail and at other times due to the fact that the model is too detailed, such that application of it in addressing a system on a particular scale is rendered intractable. The correct choice of representation is thus ultimately dependent on the spatio-temporal boundaries of the system as well as the complexities of its interactions.

Classical molecular dynamics (MD) is a deterministic method for following the trajectory of a collection of particles whose interactions are essentially described by classical dynamics and Newton's laws of motion. It is thus a method for the description of matter in its molecular representation. It can be implemented by various computational algorithms that, using a starting set of positions and velocities, calculate the forces at an initial time. Newton's equations of motion are then integrated iteratively with small increments of time, using a finite difference method to determine the next set of positions and velocities. This iterative method maps out a trajectory and thus the time evolution for the set of particles being studied.

Molecular dynamics can be used to study a range of physical problems and to measure several properties of a system. For example, it has been used to study the dynamics of fluids, of organic molecules and large biopolymers such as proteins and nucleic acids. As will be shown later, it can be used to calculate the binding affinities between proteins and inhibitors as well as being implemented to study the condensed matter physics of large systems such as clays and salts. It can also be used to calculate thermodynamic properties such as pressure and temperature and to minimise the energy of a system or to achieve thermodynamic equilibrium within a system. We will see later how MD can also be used to generate statistical mechanical data and thus describe macroscopic properties of a system from ensembles of its microscopic states.

Understanding the dynamics of proteins is essential to properly understanding their function, and here, molecular dynamics establishes a natural way of studying protein dynamics. MD has been applied to a variety of protein dynamics problems. For example, it has been used in the study of aquaporins, which are transmembrane proteins that selectively allow water molecules through the cell membrane. It has been used to demonstrate the subtleties of the highly selective process by which water diffusion is maintained through the aquaporin water channel whilst transfer of protons is prohibited [18]. Unfortunately, the large computational cost of MD largely makes it unsuitable for determining protein structure and determining the full extent of protein folding from initially unfolded states, which occurs on the μs to ms timescale, although it has been applied to the study of the refolding characteristics of small unfolded proteins [19].

In Chapter 3 we will provide examples of how MD has been applied to the study of the flexibility and dynamics of HIV-1 protease as well as the determination of drug binding affinities. In this chapter we give a short account of the theory of classical molecular dynamics, its computational implementation and some of the calculations that can be made using it. MD is very computationally intensive for a number of reasons. We will therefore look at methods that have been developed to increase the computational efficiency of MD codes as well as how various implementations afford computational enhancement through parallel processing of MD codes and how use of high performance computing (HPC) resources and grid technology allow simulations of greater size and duration to be achieved. We will end the chapter by outlining some alternative computational methods for studying biomolecules and by exploring how molecular dynamics can be coupled to various different modelling paradigms to achieve an enhanced description of biomolecular systems. There are several texts on molecular dynamics, some of which are recommended for a more thorough treatment of the subject [20–22].

2.1 Theoretical and Computational Aspects of Classical Molecular Dynamics

The theory of classical dynamics provides the description of the time evolution of a mechanical system, given the specification of the coordinates and velocities of the constituents of the system and a description for the interaction between these constituents. Given such information about a mechanical system, both the path of the constituents of that system and thus the mechanical state of the system as a whole thereafter, are determinable and can be completely described by a set of equations that interrelate the corresponding positions, velocities and accelerations. These are known as the equations of motion for the system. For a full treatment of the theory of classical dynamics from first principles, the reader is referred to the literature [23]. Here we will expound the theory in as much as is necessary for the purposes of the research performed and reported in this thesis.

2.1.1 The Equations of Motion of Many-Particle Systems

For a system containing N particles, the mechanical state of the system can be completely described, given 3N generalised coordinates q_i (where $i = 1,2,3...3N$) and 3N generalised velocities $\dot{q}_i$ alongside a potential energy function $V(q_i)$, that describes the interactions between the particles, and is dependent only on the coordinates, q_i. There are two principal formulations of the equations of motion that describe such a system. The first set of equations of motion can be derived from Lagrange's equations, which are given by:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = 0 \tag{2.1}$$

where the Lagrangian function $L(q_i, \dot{q}_i)$ is defined by the kinetic $K(\dot{q}_i)$ and potential $V(q_i)$ energies of the system:

$$L(q_i, \dot{q}_i) = K(\dot{q}_i) - V(q_i) \tag{2.2}$$

In Cartesian coordinates, where $\mathbf{r}_j$, $\dot{\mathbf{r}}_j$ $\ddot{\mathbf{r}}_j$ represent the position, velocity and acceleration vectors respectively of the j^{th} particle of mass m_j, the kinetic energy of a system is defined by:

$$K = \frac{1}{2}\sum_{j=1}^{N} m_j \dot{\mathbf{r}}_j^2 \tag{2.3}$$

Substituting Equations 2.2 and 2.3 into Lagrange's equations we are thus able to derive 3N second-order differential equations:

$$m_j \ddot{\mathbf{r}}_j = \mathbf{f}_j \tag{2.4}$$

These are Newton's equations of motion where the right hand side is termed the force $\mathbf{f}_j$ on the j^{th} particle and is given by the spatial derivative of the potential function:

$$\mathbf{f}_j = -\nabla_{\mathbf{r}_j} V \tag{2.5}$$

An alternative and equivalent set of equations of motion can be derived from the transformation of the generalised coordinates and velocities (q_i, $\dot{q}_i$) describing a system into a set of independent variables consisting of coordinates and momenta (q_i, p_i). The Hamiltonian of a system is the total energy of the system and is a function of p_i and q_i:

$$H(p_i, q_i) = \sum_{i=1}^{3N} p_i \dot{q}_i - L \tag{2.6}$$

The total differential of the Lagrangian can be expressed in terms of the Hamiltonian (for a complete derivation see the literature [23]) such that:

$$dH = -\sum_{i=1}^{3N} \dot{p}_i dq_i + \sum_{i=1}^{3N} \dot{q}_i dp_i \tag{2.7}$$

This gives rise to Hamilton's equations of motion in the independent variables p_i and q_i:

$$\dot{q}_i = \frac{\partial H}{\partial p_i} \tag{2.8}$$

$$\dot{p}_i = -\frac{\partial H}{\partial q_i} \tag{2.9}$$

Hamilton's equations are a set of 6N first order equations, as opposed to the 3N second order equations that emerge from the Lagrangian derivation. Both sets of equations equivalently describe the time evolution of a many-body system and as such form the basis of the theory of classical molecular dynamics.

2.1.2 Forcefields and the Determination of the Intermolecular Potential

Whilst the equations of motion describing a system determine the positions and velocities of the components of that system with respect to time, they are dependent on the analytical potential function V, which describes the nature of the interaction between the particles. The time evolution of a system can therefore only be determined once the form of this potential function is described.

Let us consider the potential of a system containing N atoms. It can be written as a function of the coordinates of individual atoms, pairs, triplets and so on:

$$V = \sum_i V_1(\mathbf{r}_i) + \sum_i \sum_{j>i} V_2(\mathbf{r}_i, \mathbf{r}_j) + \sum_i \sum_{j>i} \sum_{k>j>i} V_3(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) + \cdots \tag{2.10}$$

where the terms on the right hand side are component potentials from an external field, the pairwise interactions, triplet interactions etc. respectively [20]. It is important to note that the double summation accounts for all possible pairwise interactions and keeping j greater than i rules out any duplication of interactions. In fact, for a polyatomic system, the effect of all higher potential components describing the many-body effects is usually incorporated into the pairwise potential by writing down an 'effective' pair potential that usually describes the system well:

$$V = \sum_i \sum_{j>i} V_2^{eff}(\mathbf{r}_{ij}) \tag{2.11}$$

In the computational implementation of molecular dynamics, this analytical potential energy function is termed the *forcefield* and is crucial for the correct description of a polyatomic system, described by the computational program. However, description of molecular and macro-molecular systems is more complex than that of single atom systems and requires the componentisation of the potential energy beyond just pairwise terms.

There is no unique way to specify the functional form or the parametric values for such forcefields and this is evident from the differences observed in implementations developed by various groups. The AMBER forcefield [24, 25] is slightly different to either the CHARMM [26] or the GROMOS [27] forcefields. We will not discuss the variation in forcefields here, although it is important to appreciate that such variations give rise to emergent differences in studies of the same system with slightly different forcefields. As the studies conducted in this thesis have used the AMBER forcefield, we will describe the analytical potential function described by it.

Forcefield componentisation begins with the separation of bonded, V_b, and non-bonded, V_{nb}, terms:

$$V_{total} = V_{nb} + V_b \tag{2.12}$$

The bonded term accounts for contributions to the potential from bond vibrations, bond bending and torsional motions. The functional form is based on Hooke's law and the deviations from equilibrium values:

$$V_b = \sum_{bonds} K_r(r - r_{eq})^2 + \sum_{angles} K_\theta(\theta - \theta_{eq})^2 + \sum_{dihedrals} \frac{V_n}{2}[1 - cos(n\phi - \gamma)] \tag{2.13}$$

where the terms r_{eq}, θ_{eq} and γ represent equilibrium inter-atomic bond lengths, angles and dihedral angles respectively and K_r, K_θ and V_n are parameters that correspond to the stiffness of the corresponding bonded interactions. The non-bonded terms comprise the long range Coulombic electrostatics interactions V_{ele} as well as the van der Waals interactions V_{vdW} which in this case use the form of the Lennard-Jones 6-12 potential:

$$V_{nb} = \underbrace{\sum_{i<j}\left[\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}}\right]}_{V_{vdW}} + \underbrace{\sum_{i<j}\frac{q_i q_j}{\varepsilon r_{ij}}}_{V_{ele}} \tag{2.14}$$

where A_{ij} and B_{ij} are parameters that govern the repulsive and attractive components of the van der Waals interaction, ε is the permittivity of free space and q_i and q_j are the atom-centred charges. The parameters $K_r, r_{eq}, K_\theta, \theta_{eq}, V_n, \gamma, A_{ij}$ and B_{ij} are all empirically determined by grouping different combinations of atom types together and then fitting their values from experiment or from *ab initio* quantum calculations.

Transferability is a key feature of MD in that experimentally derived forcefield values on smaller systems are then used for the description of larger molecular structures. Alongside differences in functional form, it is also here that various forcefields have been parametrised differently on different sets of organic structures and usually at standard temperatures and pressures. Transferability therefore has both advantages and disadvantages. It allows for the relatively accurate treatment of larger structures such as proteins around equilibrium as the parameters themselves are validated for equilibrium structures over short timescales.

Equilibrium itself is a dynamic process and the validity of forcefields used to model systems that move away from equilibrium therefore becomes questionable and can lead to problems. Proteins can spend time in non-equilibrium conformations moving from one conformation to another and it is not certain whether MD forcefields describe some of this behaviour reliably or whether they are accurate under conditions of large temperature or pressure deviation from those under which they are derived. Another disadvantage is the limited number of multiple atomic configurations to have been parametrised. The atomic configurations of all amino acids have been parametrised, but this is not necessarily the case for the vast majority of molecules that could exist. For example, the particular atomic configuration of a novel drug may not be described with any current forcefield. The General AMBER Force Field (GAFF) tries to address this problem and extends the parametrisation to nearly all the known atomic configurations possible by organic compounds, often using parameters derived from *ab initio* methods (see § 2.7.3) [28].

Given the parametrisation of the forcefield and the positions of the atoms in a molecular structure, further computation of the force on each atom is then possible and in turn the accelerations on each of the atoms by Equations 2.4 and 2.5. The force calculation is an essential part of a molecular dynamics program. However, it alone is insufficient to determine the time evolution of the system and further requires the iterative integration of Newton's equations of motion. We will discuss some of the schemes which implement integration of these equations in § 2.1.3.

2.1.3 Integrating the Equations of Motion

In general, there is no analytical solution to the classical equations of motion of a multi-particle system interacting under a continuous potential for which the number of particles exceeds two. This is referred to as the well known 'N-body problem' in classical mechanics. A determination of the time evolution of a multi-particle system may however be achieved by numerical 'finite difference methods', which iteratively integrate Newton's equations of motion in multiple stages of a small time step δt.

In principle, finite difference methods use the positions $\mathbf{r_j}(t)$, velocities $\mathbf{v_j}(t)$ and accelerations $\mathbf{a_j}(t)$ on every atom in a system at time t, to extrapolate the atomic positions and velocities at a time $t + \delta t$. The accelerations are calculated as discussed in § 2.1.2 and it is assumed that the forces stay constant during the time step δt. Using the new positions, the new forces are calculated and the integration step is repeated to calculate new positions and velocities at a time $t + 2\delta t$ and multiple iteration thus yields the computed trajectory of the multi-particle system.

Several algorithms have been developed that all use a slightly different approach to implement finite difference method integrations of the equations of motion. However, they all begin with the assumption that the positions, velocities and accelerations at time $t + \delta t$ can be approximated by Taylor expansions:

$$\mathbf{r}(t+\delta t) = \mathbf{r}(t) + \mathbf{v}(t)\delta t + \frac{1}{2}\mathbf{a}(t)\delta t^2 + \frac{1}{3!}\mathbf{b}(t)\delta t^3 + \cdots \tag{2.15}$$

$$\mathbf{v}(t+\delta t) = \mathbf{v}(t) + \mathbf{a}(t)\delta t + \frac{1}{2}\mathbf{b}(t)\delta t^2 + \frac{1}{3!}\mathbf{c}(t)\delta t^3 + \cdots \tag{2.16}$$

$$\mathbf{a}(t+\delta t) = \mathbf{a}(t) + \mathbf{b}(t)\delta t + \frac{1}{2}\mathbf{c}(t)\delta t^2 + \cdots \tag{2.17}$$

where $\mathbf{b}(t)$ and $\mathbf{c}(t)$ are the third and fourth derivatives of the position vectors $\mathbf{r}(t)$ and $\mathbf{v}(t)$ and $\mathbf{a}(t)$, the velocity and acceleration vectors respectively. We will now outline some of the algorithms that exist for the extrapolation of the positions and velocities at $t + \delta t$, bearing in mind that a good algorithm should be computationally efficient, conserve energy and momentum, be accurate over long time steps and should duplicate the classical trajectory as closely as possible [20].

Verlet Algorithm

The method initially adopted by Verlet [29] in 1967 is, perhaps, the most widely used. It adds the Taylor expansion of both forward and reverse time steps as shown in Equation 2.18:

$$r(t-\delta t) = r(t) - v(t)\delta t + \frac{a(t)}{2}\delta t^2 - \frac{\delta t^3}{3!}\frac{d^3r}{dt^2} + \cdots \tag{2.18}$$

to give the advanced positions at $r(t+\delta t)$ and the current velocity $v(t)$:

$$r(t+\delta t) = 2r(t) - r(t-\delta t) + a(t)\delta t^2 + \cdots \tag{2.19}$$

$$v(t) = \frac{r(t+\delta t) - r(t-\delta t)}{2\delta t} \tag{2.20}$$

It can be seen from this that the positions are correct to δt^4 and the velocities to δt^2 where the first and third order terms cancel upon summing. The algorithm is also properly centred and therefore time-reversible and has been shown to conserve energy very well even at long timesteps [29]. The only problem is that the form of the algorithm introduces numerical errors because, as seen in Equation 2.19, a small term δt^2 is added to the difference of two large terms δt^0.

Modifications of the Verlet Algorithm

The deficiencies mentioned above have led to slight modifications of the Verlet algorithm such as the Leap Frog method and the Velocity Verlet as well as others. For example, the Leap Frog method [30] stores the current positions, accelerations and the previous mid-step velocities. The next mid-step velocities are calculated from the previous ones by 'leaping over' the positions which are subsequently calculated using the new mid-step velocities.

The Velocity Verlet [31] stores current positions, velocities and accelerations. It calculates the new positions and mid-step velocities. This is followed by calculating the forces half-way through the integration steps and then using the new accelerations along with mid-step velocities to calculate the new velocities. Allen and Tildesley [20] provide a very good description of several other methods including the Gear predictor-corrector algorithm.

2.2 Implementations of Molecular Dynamics

Many implementations of molecular dynamics codes, developed by different groups, exist. Examples are CHARMM [26], AMBER [32, 33], NAMD [34], GROMOS [27], GROMACS, DL-POLY [35] and LAMMPS [36]. We will not go into detail about the specific method by which each of these codes implement molecular dynamics. Instead we will briefly mention some common as well as divergent features.

Firstly, it is important to note the distinction between molecular dynamics codes and various forcefields which share common names. For example, AMBER, CHARMM and GROMOS are also names for force fields (see § 2.1.2). The AMBER package in its entirety is an ensemble of many different codes including a molecular dynamics code called SANDER, as well as protocols for building various molecular systems and for implementing several forms of computational analysis, some of which will be described in section § 2.5. NAMD is mainly a molecular dynamics code and as default uses the CHARMM forcefields. Furthermore, whilst earlier programs such as AMBER and CHARMM were written in FORTRAN, more recent programs such as NAMD, GROMACS and LAMMPS have been written either in C, or C++. One important distinction between former and latter programs is that the latter were specifically designed to utilise parallel processors more efficiently, thus affording greater speed-up over precursor serial algorithms. We will discuss the important role of parallelisation in increasing the computational efficiency of MD codes in § 2.3 as well as how different approaches are adopted by different codes. These differences subsequently lead to varying scaling capabilities when utilising larger numbers of processors.

In general, all programs require similar types of input information in order to run. Firstly, owing to the large number of conditions that can be set in a molecular dynamics simulation, all modern implementations require a configuration file, which specifies the specific conditions of the simulation. Furthermore, all implementations require coordinate information as well as information regarding the structure of the atoms within the system. Whilst coordinate files contain the positions of all atoms within the system, they contain no information about the specific structural relationships between the atoms. Such information is contained in structure files. Alongside such information it is additionally necessary to specify the forcefield.

Different programs resort to different methods for assimilating such information prior to simulation.

For example, a CHARMM forcefield consists of both a set of topology files for various amino acids, nucleotides and phospholipids as well as a parameter file containing forcefield information. A structure file for the specific system to be simulated needs to be generated from such general topology files and is usually denoted as a protein structure file (.psf file). This, together with coordinate information (.pdb file) and the corresponding parameter file, needs to be specified in the configuration file in order to correctly set up a simulation. In contrast, when using the AMBER forcefield, even though topological and parametric information is contained in separate files, such information is assimilated for a specific system into one file containing both the specific structure and parametric relationship of all the atoms in the system and is usually denoted as a .prmtop file. It is then only necessary to specify this one file alongside the coordinate information in the configuration file.

Another property of various MD codes is that although, in principle, any forcefield can be used with any molecular dynamics code, in reality, this is not immediately possible due to the discrepancies in file formatting. For example, the CHARMM parameter files have a different format to those of AMBER and whereas SANDER can readily read an AMBER forcefield file type, it cannot read the CHARMM forcefield file type. Some codes like NAMD, offer a solution whereby once system-specific structure and parameter information (.prmtop file) is generated using the AMBER package and the AMBER forcefield, it can be read by the NAMD code and thus a system can be simulated using the AMBER forcefield, but with the NAMD integrator. This is especially beneficial, given that it allows the extensive GAFF forcefield, discussed in § 2.1.2, to be used with NAMD, which offers excellent scaling and thus faster turn around, provided enough computational power is available.

2.3 Increasing Computational Efficiency

In sections § 2.1.2 and § 2.1.3, we have described the essential components of molecular dynamics (MD) and its implementation through various algorithms. In practice, MD is very computationally intensive for a number of reasons and thus various methods have been designed to produce speed up whilst retaining a sensible physical description of the system. Here we describe some of the methods used to achieve this aim.

2.3.1 Periodic Boundary Conditions

Studying the molecular properties of a system that behaves as a bulk fluid or of molecules immersed in a bulk fluid would require simulations containing a number of atoms approaching Avogadro's constant ($N_A = 6.02x10^{23}$). Simulating such a large number of atoms remains intractable and beyond the scope of computational molecular dynamics. Therefore, without any special conditions imposed upon the system, current atomistic simulations cannot reproduce the behaviour of a 'macroscopic' bulk fluid. Imposing periodic boundary conditions allows the properties of a molecular system to be studied as

though it were within a bulk fluid, whilst maintaining a relatively small size of system and allows macroscopic properties of the fluid to be calculated.

The method works by replicating the original simulation box to produce an infinite periodic array of image boxes that cover the whole of three dimensional space. The original box, therefore has a number of nearest neighbours and the coordinates of the atoms in each of the image boxes are integer multiples of the original box length added or subtracted to the original coordinates. In such a system, when an atom moves out of the original box, it is equivalent to the same atom moving into the simulation box from the image box on the opposite side. The coordinates of the atom are effectively translated to the other end of the simulation box. This conserves particle number and energy as well as allowing the atoms in the simulation to experience forces as though they were in the bulk of the fluid.

2.3.2 Constraint Dynamics

In order to correctly capture the atomic motions of molecular systems, the integration timestep necessarily needs to be relatively small in MD simulations. This is due primarily to numerical instabilities arising for larger timesteps in the integration scheme which would cause potential energies to diverge if two atoms converged on the same position. To avoid this, the timestep needs to be small enough to ensure that atomic positions can be charted before convergence of atomic position occurs and so is typically governed by the fastest oscillation of interest. As this is usually due to the X-H bond (X is any atom) [37], in small molecular systems the upper bound of the timestep is normally 1 fs. This limits the achievable simulation time for a molecular system. In biomolecular simulations, the flexibility of the X-H bond is often not of overriding importance, where instead understanding the conformational and dynamical properties of the biomolecule over longer timescales are of more interest. To facilitate such considerations, force constraints can be imposed upon a system that allow the integrator timestep to be increased, thus providing more simulation time at the expense of some accuracy. For example, making the X-H bond rigid allows the timestep to be increased to around 2 fs. An example of a constraint implementation is the SHAKE algorithm developed by Ryckaert *et al.* [38].

2.3.3 Optimising the Force Calculation

The calculation of the non-bonded component of the potential and thus the non-bonded force is the most computationally demanding part of any molecular dynamics simulation, as it requires summation of all non-bonded pairwise potentials for all atoms in the system at every time step. In order to decrease the computational demand of the required force calculation, several methods have been developed to address both the van der Waals (V_{vdW}) and Coulombic (V_{ele}) components of the non-bonded potential. The methods inherently reflect the differences in the functional forms of these two components, the van der Waals interaction being short ranged, drops off as $1/r^6$, whilst the electrostatic interaction is more

long ranged and drops off as $1/r$ (see Equation 2.14). Here, we will describe the most common method used to handle each of these forces.

Handling short-range non-bonded interactions

Due to the short ranged nature of V_{vdW}, it is possible to make the assumption that atoms far away from a particular atom under consideration make an insignificant contribution to V_{vdW} and can therefore be ignored. The computational demand of the calculation can therefore be reduced by introducing a cut-off distance, for which the V_{vdW} for all atoms separated by more than this distance is set to zero. Furthermore, by only using the nearest-neighbour periodic image boxes, V_{vdW} can be calculated for atoms that are near the boundary of the simulation box, provided that the cut-off distance is not too large that it extends beyond the image box. The cut-off should be small enough that an atom cannot interact with its own image in any of the neighbouring boxes. A cut-off distance of 10 Å is usually sufficient to model the interaction with a relatively small error [22]. There are also different schemes for assigning the functional form of the cut-off. The main schemes involve either truncating the potential at the cut-off directly, or introducing a 'switching' potential which allows V_{vdW} to smoothly tend to zero at the cut-off distance. The modification of the potential then begins at a point termed the 'switch-distance'.

Unfortunately, however, although the idea of using a cut-off is attractive, by itself there is not much decrease in computational requirement. This is due to the necessity to compute all the interatomic distances in order to determine which pairwise distances lie within the cut-off. The calculation thus becomes nearly as demanding as the calculation of the potential itself and it is therefore necessary to utilise a method that does not require the calculation of all of the distances, but which can keep track of those atoms which lie within the cut-off. The Verlet neighbour list [39] is an example of such a method. In this method, all atoms within the cut-off as well as all atoms that are within an extended spherical shell are stored in a list. As an atom's nearest neighbours do not change significantly over 10 to 20 timesteps in a fluid simulation, it is reasonable to only calculate the distances of atoms within the neighbour list and thus the contribution from atoms specified in the list that are within the cut-off over this time period. The list is then updated, keeping track of any atoms in the extended shell that have crossed into the cut-off region and vice versa as well as new atoms that have entered the extended region. Provided the extended neighbour distance is sufficiently larger than the cut-off distance, such a method ensures that atoms beyond the neighbour distance cannot enter within the cut-off before the neighbour list is updated. However, it is also important to optimise the neighbour distance and the frequency with which it is updated. If the distance is too large or the frequency of updating too great, the method becomes computationally inefficient. The converse can however, lead to errors in the energy and force calculation, as atoms may be able to enter the cut-off distance without being detected in the neighbour list first.

Handling long-range interactions

Unlike the van der Waals interaction, the $1/r$ dependence of the electrostatic interaction means that it is long-ranged and should ideally not be cut-off at an arbitrarily short distance. Atoms separated by several box lengths can still have a significant interaction with each other, and thus a desired reduction of computational demand whilst maintaining physically accurate interactions must therefore utilise other methods than those described in § 2.3.3. Several methods have been developed to handle long range interactions. Amongst them is the Ewald summation method, which is the most accurate, and will be described briefly here.

Consider the sum of the electrostatic interactions of each particle in a simulation box interacting with all other particles in the box as well as all particles in an infinite array of image boxes. This is ultimately the complete classical pairwise description of the electrostatic interaction in the MD system. However, the problem with the summation of such an interaction is that, in general, it is conditionally convergent, meaning that the sum of all positive terms and negative terms separately are naturally divergent. Furthermore, in cases where such a sum does converge to a finite value, convergence is extremely slow. An additional problem is the rapid variation of the electrostatic interaction at small distances. The Ewald method partitions the summation into two functions based on the identity:

$$\frac{1}{r} = \frac{f(r)}{r} + \frac{1-f(r)}{r} \tag{2.21}$$

where $f(r)$ is a function chosen such that both terms converge rapidly as well as being able to handle the rapid variation of the interaction at small distances. A Gaussian distribution of equal magnitude but opposite charge and centred at each atomic position, works effectively as a screening potential for each of the atomic charges. The interactions of all the atomic charges and the screening potential are then summed in real space and have the effect of diminishing the rapid changes at small separations and the property of rapid convergence. However, the introduction of the screening potential requires an exactly opposite, cancelling distribution to be summed in order to correctly calculate the Coulombic interaction. This is not efficiently summed in real space and therefore the cancelling distribution is Fourier transformed, summed in reciprocal space and the sum converted back into real space.

The reciprocal space summation is still very computationally demanding and determines the overall scalability of the force calculation, scaling as N^2, where N is the number of atoms in the simulation. Using the fast Fourier transform (FFT) method, the scaling can be improved to $NlnN$, greatly enhancing the rate of calculation of the force at each timestep. Use of the FFT method is dependent on having an ordered grid-based charge distribution, as opposed to point charges in any configuration in three-dimensional space. A variety of methods such as the particle-mesh Ewald method (PME), have therefore been developed to convert atomic charges into discretised charge distributions on a grid, where such a grid is referred to as a particle-mesh. Atomic charges are thus usually distributed amongst neighbouring

grid points on the specified particle-mesh within the simulation box, to facilitate the implementation of the FFT algorithm.

2.3.4 Parallelisation of Algorithms

The parallelisation of algorithms has been one of the most significant ways by which the turnover of MD simulations has been accelerated. The simultaneous partitioning and execution of the required MD calculations necessary for a given system across a large number of available processors (CPUs), offers huge benefits for both increasing the timescales and the size of the molecular system that can be studied in a reasonable period of wall-clock time.

However, there is no unique way to parallelise an algorithm and the variation of different implementations has subsequent consequences on the degree to which more and more processors can be used to speed up the simulation of a given system. This 'scaling' property varies across the current parallel implementations of molecular dynamics codes. In general, the greater the number of CPUs used for a computation, the greater the communication that needs to exist between such processors. As the number of CPUs increases, the increased communication load eventually results in a departure from linear scaling. For a given system size, a compromise is eventually reached between the efficiency gained by increasing the number of CPUS and the efficiency lost through increased communication loads.

Older algorithms like CHARMM and the SANDER module of AMBER, which initially used serial algorithms, have been developed further to take advantage of parallel computing environments. However, they achieve parallelisation by replicating the data of the entire system across all processors, which requires a lot of memory and results in large communication loads. Interestingly, the recently developed PMEMD module of AMBER provides better scaling than SANDER.

More recent codes such as NAMD2, GROMACS and LAMMPS decompose the calculations required for the entire system across different processors, each using different strategies. LAMMPS uses 'force decomposition', in which the pairwise forces are evenly distributed across all CPUs, whilst GROMACS uses 'particle decomposition', in which each atom is assigned to a particular CPU. NAMD2 uses 'domain decomposition', in which the system is divided into a number of spatial regions whose size is larger than the non-bonded cut-off distance, in addition to grouping together the non-bonded interactions between these spatial regions into distinct 'patches', which are also distributed across the available CPUs. The computational load across the CPUs can then be balanced at regular intervals to maximise the efficiency of each processor. This strategy is particularly advantageous for larger systems as it provides excellent scaling capabilities. The departure from efficient scaling is then largely dependent on the size of the system so that, for a large enough system, the calculation can be scaled up efficiently to thousands of processors. The turn around of a simulation is then governed by processor speed instead of system size, given enough CPUs.

For smaller systems however, it is not necessarily the scaling property that is of vital importance

but the intrinsic speed of the algorithm. For example, GROMACS is very fast and for a given system size will outperform NAMD and PMEMD at least up to 32 processors. Recent studies which have benchmarked the scaling properties of several codes such as both the SANDER and PMEMD modules in AMBER as well NAMD2, LAMMPS and GROMACS can be found in the literature [40].

2.4 The Relationship between Molecular Dynamics and Statistical Thermodynamics

One of the principal benefits of molecular dynamics methods is the ability to arrive at a statistical thermodynamical description of a system from information about the position and velocities of the constituent particles of that system. For a full treatment of statistical thermodynamics, the reader is referred to standard texts [41]. Here, we will briefly discuss the basis of such a relationship by introducing some concepts from dynamical systems theory, including the notion of *phase space*, the divergence of dynamical systems and the conditional equivalence of dynamical and statistical thermodynamical properties.

2.4.1 Phase Space and Liouville's Theorem

An N-body system moving in three spatial dimensions has associated with it 3N position vectors $\mathbf{q}$ and 3N momenta $\mathbf{p}$, corresponding to each particle. We may envisage a 6N dimensional landscape, each dimension representing one of the three-dimensionally resolved coordinates or momenta of each particle, in which each point represents a unique state of the entire system. This space is known as *phase space* and each point is a *microstate* of the system. The value of any function that is therefore dependent on the position and momenta, such as the free energy of the system, can therefore now be envisaged as forming a complex topological landscape in phase space. The time evolution of a system corresponds to a change in the unique position of the system and running molecular dynamics on the system thus allows it to explore regions of phase space such that a trajectory is charted out.

The thermodynamic properties of a system, such as temperature, pressure and volume, which are regarded as macroscopic quantities, are discernible from a statistical treatment of the possible microstates of the system. From the perspective of statistical thermodynamics, such a change in the position in phase space corresponds to a change from one microstate of a system to another. Any given thermodynamic *macrostate* of a system is characterised by many corresponding microstates available to the system. At equilibrium, the number of available microstates is maximised and these correspond to the equilibrium macrostate adopted by the system. The system continually changes from one microstate to another under time evolution. Furthermore, even though the instantaneous value of the macrostate, which is determined by the particular microstate accessed, will vary, the long time average of instantaneous values of the macrostate will be invariant at equilibrium. For example, if a gas attains an equilibrium temper-

ature, the particles of the gas may interchange kinetic energy, thus moving around in phase space and accessing different microstates, but the overall temperature, as calculated by the long time average of the instantaneous temperature, will remain constant.

Alternatively, a description of such averaged behaviour of microstates relating to a system may also be attained if one envisages a large collection or *ensemble* of systems. Each of the systems share the same macrostate but, at any moment in time, are described by differing microstates, which are in turn subject to change under time evolution. This ensemble of systems, each with varying microstates can be thought to occupy a volume in the accessible phase space and it is then possible to characterise the ensemble by a density function, $\rho(\mathbf{q},\mathbf{p},t)$ in the phase space. This density function represents the distribution of the members of the ensemble over all possible microstates at different instants of time.

Liouville's theorem describes the movement of the representative points of each system in the phase space. By considering the continuity of the ensemble of representative points in terms of the net flux of points out of and into a given volume of the accessible phase space, it states that the local density of representative points is invariant under time evolution. This is entirely similar to the manner in which an incompressible fluid moves in physical space and is given by:

$$\frac{d\rho}{dt} = \frac{\partial \rho}{\partial t} + \sum_{i=1}^{3N} \left(\frac{\partial \rho}{\partial q_i} \dot{q}_i + \frac{\partial \rho}{\partial p_i} \dot{p}_i \right) = 0 \tag{2.22}$$

where q_i and p_i are the dimensions relating to the positions and momenta of each particle in the phase space. Importantly, Liouville's theorem is not constrained to just equilibrium systems and emerges simply from considerations about the mechanical nature of the system. A derivation of Liouville's theorem can be found in texts on statistical mechanics [42].

Equilibrium however, is characterised by a stationary ensemble, one for which the density function is explicitly invariant with time:

$$\frac{\partial \rho}{\partial t} = 0 \tag{2.23}$$

This additional criterion requires that:

$$\sum_{i=1}^{3N} \left(\frac{\partial \rho}{\partial q_i} \dot{q}_i + \frac{\partial \rho}{\partial p_i} \dot{p}_i \right) = 0 \tag{2.24}$$

and thus sets particular bounds for the functional form that the density function can adopt. Furthermore, as there are a finite number of microstates describing equilibrium, the invariance of density leads to the invariance of the volume in phase space occupied by the ensemble under time evolution. Importantly, it is the number of ways with which Equation 2.24 can be solved that gives rise to a variety of different thermodynamic ensembles, which are defined by the macroscopic properties that are held constant at equilibrium.

2.4.2 Thermodynamic Ensembles and the Ergodic Theorem

The form of the density function $\rho(\mathbf{q},\mathbf{p})$ is essential in order to determine macroscopic thermodynamic properties of a system. These are evaluated by calculating the ensemble average, $\langle f \rangle$, of the desired thermodynamic quantity, $f(\mathbf{q},\mathbf{p})$ at equilibrium and given by:

$$\langle f \rangle = \frac{\int\int f(\mathbf{q},\mathbf{p})\rho(\mathbf{q},\mathbf{p})d\mathbf{q}d\mathbf{p}}{\int\int \rho(\mathbf{q},\mathbf{p},t)d\mathbf{q}d\mathbf{p}} \tag{2.25}$$

The normalised density function is thus equivalent to the probability distribution function of the microstates of the system. At thermodynamic equilibrium, the ensemble of microstates that represent equilibrium depend on the macroscopic properties of the system that remain constant, such as the number of particles N, the volume V, the energy E, the pressure P, the temperature T or the chemical potential μ. This leads to several well defined thermodynamic ensembles, namely the microcanonical (NVE), canonical (NVT), isothermal-isobaric (NPT) and the grand canonical (μVT) ensembles. For example, in the NVE ensemble, the number of particles, the volume and the energy of the system remain constant. Each of these ensembles results in a different functional form to the density function characterising the ensemble and it is through the utilisation of the correct form of the density function corresponding to each ensemble that the thermodynamic properties of that ensemble can be arrived at.

For example, in the canonical ensemble, $\rho(\mathbf{q},\mathbf{p})$ can be shown to be explicitly dependent on the Hamiltonian of the system $H(\mathbf{q},\mathbf{p})$:

$$\rho(\mathbf{q},\mathbf{p}) \propto e^{-\beta H(\mathbf{q},\mathbf{p})} \tag{2.26}$$

where $\beta = 1/k_B T$, k_B is the Boltzmann constant and T is the temperature. This form of the density function satisfies the conditions imposed by Liouville's theorem at equilibrium and weights the probability of being in a particular microstate, whereby the lower the energy of a microstate, the more probable its occurrence. Equation 2.25 can therefore be written in terms of the explicit functional form of the density function:

$$\langle f \rangle = \frac{\int\int f(\mathbf{q},\mathbf{p})e^{-\beta H(\mathbf{q},\mathbf{p})}d\mathbf{q}d\mathbf{p}}{\int\int e^{-\beta H(\mathbf{q},\mathbf{p})}d\mathbf{q}d\mathbf{p}} \tag{2.27}$$

The denominator is a normalising factor which sums over all the microstates accessible to the system in phase space. It is known as the partition function, Q_{NVT}, and can be written for the canonical ensemble as:

$$Q_{NVT} = \frac{1}{N!h^{3N}}\int\int e^{-\beta H(\mathbf{q},\mathbf{p})}d\mathbf{q}d\mathbf{p} \tag{2.28}$$

where h is Planck's constant and the prefactor takes into account the number of distinguishable microstates that exist within a given volume element, $d\mathbf{q}d\mathbf{p}$, of phase space. As the partition function is

the integral of the density function over all of the accessible microstates in phase space, its functional forms also depends on the thermodynamic ensemble adopted. In the isothermal-isobaric ensemble, the partition function, Q_{NPT}, is written as:

$$Q_{NPT} = \frac{1}{N!h^{3N}} \int\int e^{-\beta(H(\mathbf{q},\mathbf{p})+PV)} d\mathbf{q}d\mathbf{p} \tag{2.29}$$

where there is an extra PV term in the density function. Many thermodynamic properties, including the free energy of a system can be expressed in terms of the partition function relevant to that particular ensemble. Thus, knowledge of the partition function is key in order to determine the macroscopic properties of a system from their microscopic properties. In § 2.5.2 we will see how this applies to the free energy of a system in the canonical ensemble.

A real thermodynamic system does not contain a large array of ensembles, but is composed of a single system that evolves from one microstate to another in time. As the density function is equivalent to a probability function, given a suitable length of time, the time average of a single evolving system will correlate to an ensemble average of many copies of the system. However, only if the ergodic theorem is satisfied, is this the case. This states that for a system in equilibrium, the time average of a quantity is equal to the ensemble average as long as the system visits all the points it can occupy in phase space in a finite time. Furthermore, it is important to note that experimentally, when a thermodynamic property is measured, it is the long time average of the property for a single system which is being measured, even though the theoretical description of such a property is provided by an ensemble average.

It is this equivalence between an ensemble average of a physical quantity and the long time average of the same quantity that enable the thermodynamic quantities of a system at equilibrium to be calculated using molecular dynamics. Generation of molecular dynamics trajectories at equilibrium is thus equivalent to charting the time evolution of the microstates of the system, the average of which should yield a correct thermodynamic ensemble average for a given property. An inherent problem with generating such temporal snapshots is that the time average of such snapshots, for a given property, is not necessarily representative of a thermodynamic ensemble average, due to insufficient temporal sampling. Fulfilling the ergodic theorem, sampling for long enough and ensuring that the molecular system is set up in a way that corresponds to the required ensemble of the thermodynamic measurement being taken, allows a whole host of thermodynamic properties to be determined about a system.

For example, for a system in the microcanonical ensemble, the instantaneous temperature T of a system can be calculated via the equipartition theory for a system with n degrees of freedom, from the average kinetic energy of the particles at any one instant of time:

$$\frac{n}{2}k_B T = \left\langle \frac{1}{2}mv^2 \right\rangle \tag{2.30}$$

where m is the mass of a particle and v, its velocity. The ensemble average of the temperature $\langle T \rangle$ as

calculated from a time average of the values of instantaneous temperature would then yield the correct thermodynamic temperature at equilibrium. Equation 2.30 can be derived by considering the probability distribution $f(v)$ for the velocities v of the constituent particles of a system with mass m at a specified temperature T. Even though the velocities of particles in a system can be vastly different from each other, they are described well by this probability distribution, known as the Maxwell-Boltzmann distribution:

$$f(v) = 4\pi\left(\frac{m}{2\pi k_b T}\right)^{3/2} v^2 exp\left(\frac{-mv^2}{2k_b T}\right) \tag{2.31}$$

The probability of finding a particle with a velocity of magnitude between v and $v+dv$ is then $f(v)dv$ and the integration of this velocity distribution multiplied by the kinetic energy of a particle, from zero to infinity, yields the average kinetic energy of each particle. The Maxwell-Boltzmann distribution is important in molecular dynamics simulations as it is used to assign the velocities of every atom in a system when only the required temperature of the system is specified.

Alongside the simple thermodynamic properties considered here, it becomes possible to calculate properties more pertinent to biomolecular systems, such as the free energies of ligand binding as well as the conformational and dynamical properties of biomolecules at equilibrium. These will be discussed further in § 2.5.

2.4.3 Maintaining the Thermodynamic Ensemble

Molecular dynamics simulations evolving under an integration scheme that conserves energy and under periodic boundary conditions sample from the microcanonical (NVE) ensemble, due to an implicit conservation of particle number, volume and total energy. The construction and maintenance of a system in alternative thermodynamic ensembles requires the implementation of additional protocols in molecular dynamics simulations. Classical MD is currently capable of reproducing the NPE, NVT and NPT ensembles alongside the NVE and does this using the employment of certain thermostats and barostats. These are essentially methods developed to constrain either the temperature of a system or its pressure. Whilst we will briefly mention some of these methods, for a more detailed exposition the reader is referred to the literature [20].

Maintaining a constant temperature can be achieved in a number of ways. One method involves rescaling the velocities of each particle by a factor which restores the temperature to the desired value, whilst maintaining the Maxwell-Boltzmann distribution. The Anderson thermostat [43] implements such a method and keeps the kinetic energy constant but does not maintain a canonical (NVT) ensemble. Another approach is to treat the system as though it were in thermal contact with an external 'heat reservoir'. Energy exchange is permitted both ways between the reservoir and the system and the reservoir is described by an additional degree of freedom in the system. An example of this is the

Nosé-Hoover thermostat [44–46] which conforms accurately to the canonical ensemble and allows a gradual return to a desired temperature as opposed to absolute constancy of a specified temperature. Langevin dynamics [47, 48] can also be used to maintain the temperature. In this approach the thermal properties of an external bath at the boundaries of the system are modelled in such a way that particles near the boundary of the system feel both a frictional retardation as well as a random force from the bath representing high energy collisions from 'bath' particles. This is implemented by introducing a frictional term with coefficient γ_i and a stochastic force $\mathbf{R}_i$ into the equations of motion for each particle i, the coupled interaction of which allows the system to tend to a target temperature.

$$m_i\ddot{\mathbf{r}}_i(t) = \mathbf{F}_i(t) - \gamma_i m_i\dot{\mathbf{r}}_i(t) + \mathbf{R}_i(t) \tag{2.32}$$

Here, m_i is the mass of particle i, $\mathbf{F}_i$ is the normal interatomic force on the particle and $\mathbf{R}_i$ is applied in such a way that the mean stochastic force is zero. The Langevin approach is thus also an approximate description of Brownian motion, in which the random motions of particles suspended in a fluid are determined by the bombardment of the molecules of the fluid.

Furthermore, the validity of the Langevin approach in restoring the temperature to a designated value whilst maintaining equilibrium is provided through the fluctuation dissipation theorem. This states that at equilibrium, the response of a system to a small external perturbation is the same as its response to a spontaneous fluctuation. This correlates the frictional coefficient γ_i to the magnitude of the applied stochastic force such that different values of γ_i will cause the system to return to an equilibrium temperature at different rates. The value of γ_i adopted, therefore, has to be optimised between a large value, for which stochastic perturbations which increase temperature would dissipate quickly, and a small value, for which stochastic perturbations would dominate the relaxation of a system, below the required temperature, back to its desired value.

Pressure is controlled using similar methods to that of temperature. One method involves coupling the system to an external 'piston' which responds to the force of the system and changes the volume of the box to match the required pressure [43]. Another method is that implemented by the Berendsen barostat [49]. In this approach, the system is coupled to a pressure bath by introducing an extra term into the equations of motion such that the instantaneous pressure $P(t)$ of the system tends to the desired pressure of the surrounding bath P_{bath} through a rescaling of the volume of the box. The rate of change of instantaneous pressure is given by:

$$\frac{dP(t)}{dt} = \frac{1}{\tau_p}(P_{bath} - P(t)) \tag{2.33}$$

where τ_p is the pressure coupling parameter. The larger this parameter, the more rapidly the instantaneous pressure approaches the bath pressure and the tighter the coupling between the system and the bath.

2.5 Calculable Properties in Molecular Dynamics

Molecular dynamics can be used to calculate a whole range of dynamical and thermodynamical properties of a system. In liquids, these can be properties such as diffusion constants and transport coefficients as well as shear, heat flow and Brownian dynamics [20]. It can also be used to calculate important properties of biomolecules such as conformational flexibility, hydrophobic and hydrophilic interactions between proteins, solubility and the strength of protein-ligand binding amongst others, providing insight into a whole range of biomolecular processes. However, we will not provide an exhaustive account of all the properties of liquids and solute systems that can be calculated using methods developed to utilise molecular simulation, as this is beyond the remit of this thesis. For this the reader is referred to standard texts [20, 21]. Instead, we will focus our discussion around the analytical methods used to provide insight into the studies conducted in this thesis. These insights fall into two broadly defined categories, termed 'qualitative' and 'quantitative', which will be discussed further here.

2.5.1 Properties that Provide Qualitative Insight

Broadly speaking, 'qualitative' insight refers to insight gained from analytical methods that provide a non-thermodynamic description of the system. Such methods take advantage of the time evolution of a dynamical system to gather information about conformational properties of a protein as well as flexibility, fluctuations and dynamics of parts of proteins which may result in insight into protein function that are difficult to ascertain through experimental means.

Crucial to such insight is the availability of good visualisation packages for molecular systems. Visualisation of a simulation enhances insight by allowing the scientist to 'see' much more clearly what a system is doing and from this attempt to gain more quantifiable information about certain properties of interest. Visual Molecular Dynamics (VMD) [50] is an example of an excellent visualisation package which we have used to generate several figures in subsequent chapters of this thesis.

Following visualisation, which is often a primary means of analysing a simulation, many methods exist for quantifying protein dynamics. We will discuss the basis of three analytical methods here in order to facilitate the application of such methods in later chapters.

Root Mean Squared Deviation (RMSD) Analysis

RMSD analysis is one of the principal methods used to calculate biomolecular flexibility and conformational changes. The RMSD of a set of atoms in a molecular simulation measures the average deviation of those atoms from the same atoms in a pre-specified reference system. As a molecular simulation evolves, the coordinates of the system vary from the original structure and their RMSD can be calculated relative to the original starting structure by:

$$R(t) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mathbf{r}_i(t) - \mathbf{r}_i(0))^2} \tag{2.34}$$

where N is the number of atoms being considered, $\mathbf{r}_i(t)$, the position of atom i at time t and $\mathbf{r}_i(0)$, the original starting position of atom i.

However, without prior alignment of a set of atoms, RMSD analysis does not measure the conformational changes in a molecular structure. Consider a protein moving only in a rotational and translational sense. Whilst time evolution of the RMSD of such a system will increase with respect to the starting structure, the tertiary structure of the protein may not have changed. An RMSD measurement will in this case provide the total deviation of the structure from its original position. However, if the coordinates of the protein atoms are aligned at each snapshot in the trajectory, then the RMSD will give a measure of the conformational deviation of the protein.

RMSD analysis is used as an indicator of equilibration in a molecular system. Whilst thermodynamic equilibrium may be reached in a molecular simulation of a protein, increase in the RMSD of the protein indicates that conformational readjustments from the starting structure are still occurring. Equilibration may be marked typically by a 'plateau' in the RMSD with respect to the starting structure. The protein's RMSD will then fluctuate around this plateau value; the size of these fluctuations indicates the flexibility of the protein, which can be calculated using the root mean squared fluctuation (RMSF). The RMSF is the mean RMSD calculated with respect to the average structure of the protein once the plateau is reached.

However, an RMSD plateau alone is not a guaranteed measure of equilibration. True equilibrium is achieved when a system populates only it's equilibrium region of phase space. A system moving continuously away from any one region in phase space and thus not exhibiting equilibrium may also yield an RMSD plateau if it's distance in phase space is approximately maintained with respect to the initial starting structure.

It is useful to compare time-averaged RMSDs with those from an ensemble of static crystal structures. Pairwise RMSDs of crystal structures can be calculated by first aligning two structures according to a desired protocol and then measuring the RMSD using Equation 2.34, where instead of comparing atomic positions at different times, atomic positions between the two aligned structures are considered. Similarly the RMSF can be calculated by the average of the pairwise distribution of RMSDs of an ensemble of static structures. RMSDs and RMSFs are considered in Chapters 3-7 of this thesis.

Interestingly, the thermal factor (B-factor) in X-ray crystal structures (see Chapter 1) can be reproduced from atomic fluctuation data. Single atomic fluctuations are merely the RMSF of single atoms and the B-factor (B_i) of a particular atom i is given by:

$$B_i = \frac{8}{3}\pi^2 \langle \Delta r_i^2 \rangle \tag{2.35}$$

where Δr_i is the displacement of atom i. The mean squared displacement $\langle \Delta r_i^2 \rangle$ across the data set is then proportional to the B-factor.

The Radial Distribution Function (RDF)

The radial distribution function (RDF) is a pair correlation function mainly used for analysing the structure of fluids. Unlike an ideal gas, in which the atoms are randomly distributed, the interatomic interactions in most liquids lead to particular clusterings at the microscopic level. For example, non-bonded interactions between molecules induce a structure whereby the molecules on average arrange themselves at preferred distances, known as 'shells'. The RDF calculates the ratio of the average number of atoms in a spherical shell at a distance r from each atom to the number one would expect in the same shell for an ideal gas. It is given by:

$$g(r) = \frac{\langle n(r) \rangle}{\rho 4\pi r^2 \Delta r} \tag{2.36}$$

where ρ is the density of the liquid. Taking each atom in the simulation as the centre, the number of atoms in a shell between distance r and $r + \Delta r$ is calculated from which the average $\langle n(r) \rangle$ is calculated and compared to the value for an ideal gas, thus giving a dimensionless quantity $g(r)$, which is the RDF.

In the context of simulations of biomolecular structures, the RDF can be used to determine changes in the structure of water around solute molecules and thus provide information about solvation and the interaction of water with the solute. An example of this is given in Chapter 5.

Cross-Correlation Coefficients

Cross-correlation coefficients calculate the degree of correlation between two varying quantities. Like the calculation of the RMSF, either the time-averaged value of each quantity with respect to the mean is calculated or, for a static ensemble of systems, the average can be taken over the pairwise distribution of the ensemble. In biomolecular structures, it is particularly interesting to calculate positional cross-correlations between atoms or groups of atoms, in order to determine which components of a structure move together and thus elucidate conformational properties.

The positional cross-correlation coefficient between groups of atoms i and j is given by:

$$C_{ij} = \frac{\langle \Delta r_i(t) \Delta r_j(t) \rangle}{\sqrt{\langle \Delta r_i(t)^2 \rangle \langle \Delta r_j(t)^2 \rangle}} \tag{2.37}$$

where $\Delta r_i(t)$ in the time-averaged case is the displacement of the centre of group i, at time t, from the centre of the same group in the time averaged structure. Alternatively, when calculating the cross-correlation across an ensemble of static systems, $\Delta r_i(t)$ represents the displacement of atom i between the t^{th} pair of systems.

The form of the cross-correlation coefficient is normalised such that it varies between +1, representing perfect correlation, and -1, representing perfect anti-correlation, and where a value of 0 implies no correlation. Furthermore, it is useful to arrange cross-correlation coefficients into matrices corresponding to an array of atomic groups. For proteins, this could be an array of all of the amino acids of the protein. Such matrices are known as cross-correlation maps (CCMs) and are useful in providing insight into conformational properties, such as secondary structure of proteins, as well as exposing coupled interactions between residues that may be structurally distant. An example of a CCM is given in Chapter 3.

Principal Component Analysis (PCA)

In general, principal component analysis (PCA) is a technique used for the reduction of multidimensional data sets to fewer dimensions that represent most of the relevant variance within the data. Applied to the dynamics of biomolecular systems, it is a tool which can be used to reduce the number of degrees of freedom of the system so that the important modes of motion in the system can be captured. Consider the motion of a protein in water. The dynamics of the atoms of the protein consists of both random thermal fluctuations as well as concerted atomic movements corresponding to the global modes of motion of the protein. The meaningful 'slow' modes of motion are the concerted directions of atomic motion along which most of the variance in the actual dynamics is captured and can correspond to the structural flexibility of the protein at longer timescales [51], not accessible within the MD timescale itself. PCA allows these slow modes to be determined and separated from the faster modes, which capture mostly the thermal noise of the system. Furthermore, PCA allows the conformational sampling of the reduced phase space representative of these slow modes to be determined [52].

We will now outline the theoretical basis of PCA and how it is implemented in practice. Let $\mathbf{q}(t) = (q_1(t), q_2(t), q_3(t), ..., q_{3N}(t))$ be the coordinate vector of a system of N atoms in configurational space at time t. For an MD trajectory containing M temporal snapshots, the covariance matrix $\mathbf{C}$ is a $3N$-squared matrix with elements c_{ij}:

$$c_{ij} = M^{-1} \sum_{t=1}^{M} (q_i(t) - \langle q_i \rangle)(q_j(t) - \langle q_j \rangle) \tag{2.38}$$

where $\langle q_i \rangle$ is the mean value of q_i across the trajectory. The *eigenvectors* and *eigenvalues* of the covariance matrix can be determined by diagonalising $\mathbf{C}$, such that:

$$\Lambda = \mathbf{V}^T \mathbf{C} \mathbf{V} \tag{2.39}$$

where Λ is a diagonal matrix whose diagonal elements are the $3N$ eigenvalues $\lambda_1, \lambda_2, \lambda_3 ... \lambda_{3N}$, whilst $\mathbf{V}$ is a square matrix containing the corresponding $3N$ eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3 ... \mathbf{v}_{3N}$, each of dimension $3N$:

$$\mathbf{V} = \begin{pmatrix} v_1(1) & \cdots & v_{3N}(1) \\ \cdots & \cdots & \cdots \\ v_1(3N) & \cdots & v_{3N}(3N) \end{pmatrix} \tag{2.40}$$

The eigenvectors with the largest eigenvalues are called the principal component eigenvectors and represent the slow modes of motion of the system. The extent to which the trajectory conformationally samples these principal eigenvectors is given by the time evolution of the principal component projections $p_i(t)$:

$$p_i(t) = \mathbf{v}_i.(\mathbf{q}(t) - \langle \mathbf{q} \rangle) \tag{2.41}$$

where i denotes the i^{th} principal component and $\langle \mathbf{q} \rangle$ is a vector of the time-averaged value of each of the $3N$ components of $\mathbf{q}(t)$ across the trajectory.

In practice, the net rotational and translational modes of motion are usually first subtracted when analysing biomolecules using PCA [51]. This is implemented by least-squares alignment of the coordinates of the desired structure to an appropriate reference structure, such as the structure of the first snapshot in the trajectory, prior to the calculation of the covariance matrix. Alignment can then be improved by determining the average structure and realigning, using this average structure as the reference. The diagonalisation step is the most time consuming and this limits the size of the system that can be processed. PCA can be intractable for several thousand atom systems, especially if implementations of it are not parallelised. It is thus normal to include only a subset of atoms, such as only the C_α or the backbone atoms (C_α, C, N) in proteins. The PTRAJ module in the AMBER 9 software package [53] contains a serial implementation of PCA which has been used in analyses conducted in this thesis.

An interesting feature of PCA is that the eigenvectors and projections can be used to generate animations of the principal components, thus allowing visualisation of the principal modes of motion of a particular molecule. Interactive Essential Dynamics (IED) [54] is an example of a software package which does this by interfacing with the VMD visualisation package. In Chapter 5 we implement PCA to investigate the differential modes of motion of the inhibitor saquinavir in the active site of various HIV-1 protease mutants and represent our results using both the IED and VMD software packages.

Steered Molecular Dynamics (SMD)

The limiting timescales sampled using conventional MD prohibit the mechanisms of many biomolecular processes from being observed. For example, the binding and unbinding of ligands to proteins as well as the folding or unfolding of proteins are events that occur on the μs to ms timescale and conventional MD is currently unable to explore the mechanisms by which these events occur. The ability to 'steer' the components of a simulation in a desired direction can therefore provide great insight into the likely mechanisms for such biomolecular events.

Steered molecular dynamics (SMD) is a technique used to apply forced steering to an MD system [55]. It has been used to elucidate the binding/unbinding pathways of ligands in proteins such as biotin in avidin [56], retinal in bacterio-opsin [57], phosphate release in actin [58] and of anti-retroviral inhibitors binding to HIV-1 reverse transcriptase [59] as well as others.

SMD involves the selection of an atom or group of atoms (SMD atoms) to be steered in a designated direction (SMD direction). The direction can be linear or one which changes along the steering pathway. Two main methods are used for implementing SMD, 'constant velocity' or 'constant force' steering. Constant velocity steering involves attaching the centre of mass of the SMD atoms to a restraint point, known as a 'dummy atom' with a harmonic potential U, of spring constant k, and then moving the 'dummy' atom with a constant velocity v. U is then given by:

$$U = \frac{1}{2}k[vt - (\mathbf{r} - \mathbf{r}_0).\mathbf{n}_s]^2 \tag{2.42}$$

where t is the time, $\mathbf{r}$ is the position of the centre of mass of the SMD atoms, $\mathbf{r}_0$ is their original position at $t = 0$ and $\mathbf{n}_s$ is the SMD direction vector. It is then appropriate to analyse the SMD force ($\mathbf{F}$) along the steered pathway, which is given by:

$$\mathbf{F} = -\nabla U \tag{2.43}$$

It is important to select appropriate values for the parameters k and v; small values of k will cause the 'dummy' atom to move without significantly moving the SMD atoms, whilst a high value of v will cause the force profile to lack detail. Constant force steering involves applying a constant force to the centre of mass of the SMD atoms, applied in the SMD direction. It is then more appropriate to analyse the distance moved by the centre of mass during the course of the simulation. SMD has been implemented in the molecular dynamics package NAMD2 and we use this implementation in Chapter 5 to study the unbinding of the inhibitor saquinavir from the active site of HIV-1 protease. In § 2.6.2 we discuss how high performance computing confers additional benefits to SMD methods.

2.5.2 Quantitative Insight from Free Energy Calculations

Quantitative insight refers broadly to the energetic information that can be gathered about a system. This can range from the free energy of binding of biomolecules in thermodynamic equilibrium [60] to methods that calculate the energetics of particular protein conformations as well as energetic barriers along conformational pathways using a variety of sampling techniques. We will restrict our discussion to free energy methods.

As mentioned in Chapter 1, calculation of the free energy of binding is an important objective in the evaluation of the strength of protein-ligand interactions. Molecular dynamics generates a substantial

quantity of data about a system at the molecular level, which makes it theoretically a very attractive method to determine the molecular free energy of binding.

For systems in the isothermal-isobaric (NPT) ensemble, equilibrium is achieved at the minimum Gibbs free energy (see Equation 1.2). Let us recall that the stronger the enzyme-inhibitor binding, the more negative is ΔG. The free energy of binding is comprised of the enthalpic contributions from the molecular interactions of the enzyme and inhibitor, as well as the entropic contribution that arises mainly due to conformational changes of both the solute and the solvent upon binding.

In principle, it is possible to employ MD to determine the exact free energy of a system. Unfortunately, in practice, it is infeasible to calculate absolute free energies of a system using MD, as such calculations converge very slowly. We will show this by deriving the Helmholtz free energy of a system, F, in the canonical ensemble in terms of the energies of its microstates, although the same convergence problem occurs in the isothermal-isobaric ensemble, for which the Gibbs free energy function, G, is used.

The free energy of a system in the canonical ensemble is given by the Helmholtz function, $F = U - TS$, where U is the internal energy, T the temperature and S the entropy. This free energy can be written in terms of the canonical partition function, Q_{NVT}:

$$F = -\frac{1}{\beta} ln Q_{NVT}(\mathbf{q},\mathbf{p}) \tag{2.44}$$

Substituting Equation 2.28 into Equation 2.44, we have:

$$F = -\frac{1}{\beta} ln\Big(\frac{1}{N!h^{3N}} \int\int e^{-\beta H(\mathbf{q},\mathbf{p})} d\mathbf{q}d\mathbf{p}\Big) \tag{2.45}$$

Multiplying the denominator within the parentheses by unity , in the form $\int e^{-\beta H(\mathbf{q},\mathbf{p})} e^{+\beta H(\mathbf{q},\mathbf{p})} d\mathbf{q}d\mathbf{p}$, rearranging and discarding the constant $\frac{1}{N!h^{3N}}$, we have:

$$F = \frac{1}{\beta} ln\Big(\frac{\int\int e^{+\beta H(\mathbf{q},\mathbf{p})} e^{-\beta H(\mathbf{q},\mathbf{p})} d\mathbf{q}d\mathbf{p}}{\int\int e^{-\beta H(\mathbf{q},\mathbf{p})} d\mathbf{q}d\mathbf{p}}\Big) \tag{2.46}$$

From Equation 2.27 we recognise that the term in parentheses can be written in terms of an ensemble average, giving:

$$F = \frac{1}{\beta} ln\Big\langle e^{\beta H(\mathbf{q},\mathbf{p})}\Big\rangle \tag{2.47}$$

and for the Gibbs free energy, G, in the NPT ensemble (although not derived here):

$$G = -\frac{1}{\beta} ln\Big\langle e^{\beta (H(\mathbf{q},\mathbf{p})+PV)}\Big\rangle \tag{2.48}$$

The free energy is thus directly related to the ensemble average of the exponential of the Hamiltonian. Calculation of this ensemble is a thermodynamically 'exact' way of calculating the free energy of

the system. However, due to the dependence on this exponential form and not a linear function of the energy, microstates with high energies, even though sampled infrequently, contribute significantly to the calculated value. Accurate calculation of the free energy requires adequate sampling of lower energy states and it is the difficulty in ensuring that regions of phase space that make important contributions to the free energy are adequately sampled, that leads to a slow convergence time for the ensemble average. Unfortunately the convergence time for calculating ensemble averages of the entire Hamiltonian are well beyond practical means and so it is not possible to use such an approach.

Nonetheless, several 'exact' methods have been developed for the calculation of free energy differences [60], such as Free Energy Perturbation (FEP) and Thermodynamic Integration (TI). These use not the calculation of the entire Hamiltonian, but only those components that are changing, to calculate the change in the Hamiltonian ΔH between two states and this gives the free energy difference ΔG with much smaller convergence. Unfortunately, the overriding limitation of such theoretically 'exact' methods is that they are very computationally expensive.

More approximate methods also exist that rely on making assumptions about a system and some degree of empirical parametrisation. Examples are the Molecular Mechanics Poisson-Boltzmann Solvent Accessible surface area (MMPBSA) method which is much less computationally demanding than TI and FEP, and the Linear Response (LR) method, which requires fitting to experimental data, but is by far the most rapid in comparison. We will discuss some of the above-mentioned methods as well as how the construction of 'thermodynamic cycles' is employed in calculations of the free energy difference of binding.

Thermodynamic Cycles

Unfortunately, for the free energy of binding in solution, it is still impractical to calculate the ΔG of going from an unbound to a bound state as the convergence times are still too large. It is possible to employ the use of two principal thermodynamic cycles (see Figure 2.1) to calculate either the **relative free energy difference** of binding ($\Delta\Delta G$) of two sets of enzyme-inhibitor complex that are different from each other, or the **absolute free energy difference** of binding (ΔG) in solution.

The basis of all thermodynamic cycles is the fact that the free energy is a function of state, and as such, the difference in free energy between two states is independent of the route taken between them. This allows the determination of the free energy difference along a particular pathway to be determined via alternative pathways, as long as the terminal states are the same.

In the first case (see Figure 2.1(a)), the free energy difference is computed for the unphysical 'alchemical' mutation of, for example, one inhibitor to another or one enzyme into a mutant version for both the unbound (ΔG_1) and the bound (ΔG_2) systems. The difference in these two unphysical calculations are then equivalent to the differences in the physical or 'real' free energy differences of binding of the two types of enzyme-inhibitor complex:

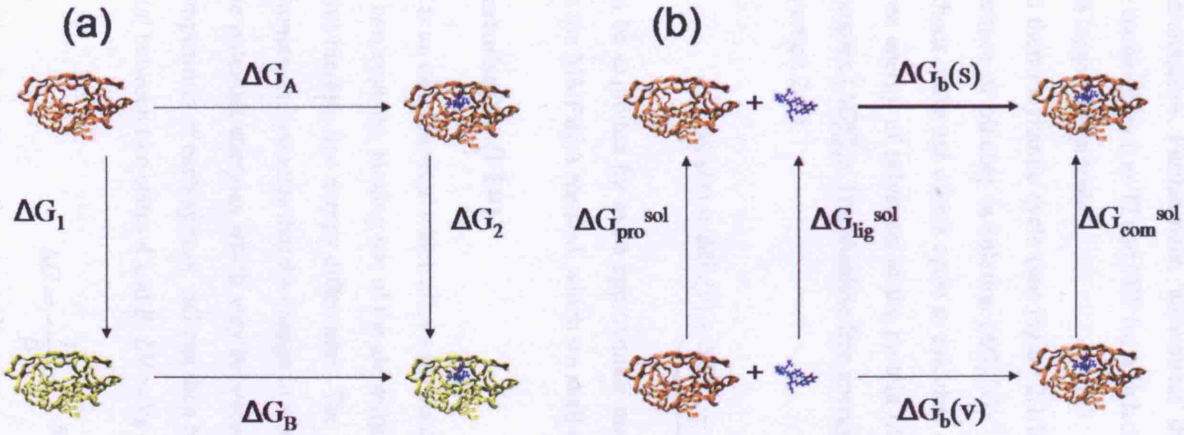


Figure 2.1: (a) Alchemical thermodynamical cycle employed to circumvent the difficulty of calculating the absolute free energy of binding (ΔG_A or ΔG_B) of a ligand to a protein. Instead the free energy of mutation of protein A (orange) to protein B (yellow) is calculated for both the unbound (ΔG_1) and bound (ΔG_2) systems, giving the relative free energy difference of binding ($\Delta\Delta G$) of the two systems (see Equation 2.49). (b) Solvation thermodynamic cycle used to calculate the absolute free energy of binding in solution ($\Delta G_b(s)$), which is intractable directly. Instead, the free energy of binding *in vacuo* ($\Delta G_b(v)$) is calculated for a ligand binding to a protein as well as the free energies of solvation of the ligand (ΔG_{lig}^{sol}), protein (ΔG_{pro}^{sol}) and the protein-ligand complex (ΔG_{com}^{sol}). $\Delta G_b(s)$ is thus arrived at indirectly through Equation 2.50.

$$\Delta\Delta G = \Delta G_B - \Delta G_A = \Delta G_2 - \Delta G_1 \tag{2.49}$$

This convenient relationship allows a comparison of the free energy differences of different sets of enzyme-inhibitor complexes, whether this be to differentiate the efficacy of one inhibitor relative to another or to determine the effects of mutation of one enzyme with respect to its wildtype in relation to specific drug interactions. Furthermore, 'alchemical' thermodynamic cycles are exploitable by more exact free energy methods such as TI and FEP for which computation of absolute binding energy differences (ΔG) is still largely intractable.

In the second thermodynamic cycle (see Figure 2.1(b)), the objective is to determine the absolute free energy difference of binding in solution ($\Delta G_b(s)$). As this is intractable via a direct approach, approximate methods make use of this cycle to calculate the free energy difference *in vacuo* ($\Delta G_b(v)$), as well as the free energy of solvation of the protein (ΔG^{sol}_{pro}) and ligand (ΔG^{sol}_{lig}) separately and the protein-ligand complex (ΔG^{sol}_{com}). The absolute free energy difference of binding in solution ($\Delta G_b(s)$) is then given by Equation 2.50:

$$\Delta G_b(s) = \Delta G_b(v) + \underbrace{\Delta G^{sol}_{com} - (\Delta G^{sol}_{lig} + \Delta G^{sol}_{com})}_{\Delta G^{sol}} \tag{2.50}$$

Such a cycle can be exploited by more approximate methods that aim to calculate absolute binding affinities, such as the MMPBSA method, which we shall discuss shortly.

Free Energy Perturbation (FEP)

The FEP method is an exact method which allows for the calculation of the 'alchemical' free energy difference between two systems, making use of the alchemical thermodynamical cycle (see Figure 2.1(a)) to calculate relative binding free energy differences. The method separates out the kinetic energy from within the Hamiltonian and assumes that the change in kinetic energy between two systems will be zero. It is then only the potential energies which vary between system A and B, owing to the differences in the molecular composition of each system. ΔG can then be expressed in terms of the ensemble average change in potential between two states A and B, $\Delta V = V_B - V_A$:

$$\Delta G = -\frac{1}{\beta} ln \left\langle e^{-\beta \Delta V} \right\rangle_A \tag{2.51}$$

where the angular brackets denote the ensemble average. Here, a molecular trajectory is charted using the potential of system A. However, a new trajectory is not obtained using V_B and instead the potential for system B is calculated using the atomic coordinates obtained for the trajectory of system A. This is denoted by the symbol A outside the angular brackets.

This requires the ensemble of atomic coordinates to be such that the potential energy V_B is also in a low energetic configuration. In practice, the method usually requires the use of several unphysical

intermediate states with an assigned progression parameter λ that gradually changes the MD potential function (from $\lambda = 0$ to $\lambda = 1$). This effectively grows the new set of atoms belonging to the end point system and diminishes the atoms belonging to the original whilst preserving the forcefield on all unchanging atoms.

Thermodynamic Integration (TI)

TI is another exact method for calculating the alchemical free energy difference and uses the integral of an ensemble average of the derivative of potential energy with respect to the same fictitious parameter λ that represents the gradual conversion of one end-point physical state to another. In TI, the free energy is expressed in terms of an integral across the parameter λ:

$$\Delta G = \int_0^1 \left\langle \frac{\partial V(\lambda,x)}{\partial \lambda} \right\rangle \partial \lambda \tag{2.52}$$

where the term in angular brackets denotes the ensemble average of the rate of change of the potential with respect to the parameter λ. The integral is effectively determined by performing numerical integration through evaluating a set of discrete simulation runs at values of λ that vary from 0 to 1 and again requires coupling λ to the MD potential function. Its advantage over FEP is that each ensemble average is calculated from its own set of atomic coordinates and is not dependent on the previous set.

In order to calculate the relative free energy differences of binding of two non-identical systems, both the FEP and TI methods require the use of the alchemical thermodynamic cycle (see Figure 2.1(a)). This requires conducting two sets of simulations, one for the alchemical mutation in the bound state and one in the unbound state. The λ parameter varies for each run within both sets of simulations and as it is necessary to vary λ in small increments, in order to sample more overlapping regions of phase space in FEP, or to gain more points for the numerical integration in TI, this means that each set can contain of the order of 20 simulations. This makes a total of around 40 simulations, each of which can be several nanoseconds long [61], due to the requirement to equilibrate the system first and the conditional convergence time for the ensemble average being calculated.

Both FEP and TI are thus very computationally intensive, requiring typically of the order of 100 nanoseconds of simulation of biomolecular systems which can contain of the order of 10^5 atoms. Unless high performance computational resources are available, the undertaking of these calculations can be intractable on realistic timescales.

The Linear Response (LR) Method

The linear response (LR) method, also known as the linear interaction energy (LIE) method is an approximate, semi-empirical method that describes the free energy of binding solely in terms of the changes in electrostatic and van der Waals interactions between the solvated ligand and the solvated protein-ligand

complex [62]. The free energy of binding is given by:

$$\Delta G = \beta\left(\left\langle V^{el}_{lig-pro}\right\rangle - \left\langle V^{el}_{lig-sol}\right\rangle\right) + \alpha\left(\left\langle V^{vdw}_{lig-pro}\right\rangle - \left\langle V^{vdw}_{lig-sol}\right\rangle\right) \tag{2.53}$$

where the parameters α and β are to be determined, the angular brackets denote ensemble averages and the *lig-pro* and *lig-sol* terms indicate solvated complex and solvated ligand systems respectively. It is therefore necessary to run only two simulations, one for the solvated complex and the other for the solvated ligand. There is no unique way to determine the parameters α and β and these are, in principle, determined empirically by comparing against a set of protein-ligand complexes of experimentally known binding free energies. Once the parameters have been set by simulating a range of protein-ligand complexes, the free energy function can be used to predict the free energies of binding of other protein-ligand complexes.

In early studies, β was set to equal 0.5, a result that emerges from dielectric theory [22], and only α was determined empirically [63, 64]. The values of α determined were consistent and such studies gave good predictions of free energies of binding for systems not included within the empirically fitted data set. Other studies, however, supported the notion that different values of α and perhaps β were required. This was not only due to the fact that different forcefields may result in a different value for α and β, but also because these parameters may depend on the nature of the binding site [65]. In some studies, a third term was introduced, to allow for a positive solvation contribution resulting from the penalty induced in forming a solute cavity. This was applied especially to problems regarding the free energy of hydration. The additional term was proportional to the solvent accessible surface area and its constant of proportionality, γ, was empirically determined alongside α and β [66].

Therefore, a range of parameter sets exist for the LR method. No one set has been successfully applied to all protein-ligand complexes and it is therefore unknown whether the best strategy is to pursue a parameter set that can do this, or to fit only a limited set of protein-ligand complexes. The latter is more beneficial for the accurate prediction of complexes not too dissimilar from those used to fit the free energy function, but would be limited if applied to a whole host of largely varying complexes. Conversely, it may be easier to predict, up to a moderate accuracy, the free energy of binding of a particular complex using a generalised function that is fitted for a whole host of largely different protein-ligand complexes, but it may not be possible to achieve high levels of accuracy with such a strategy.

The Molecular Mechanics Poisson-Boltzmann Solvent Accessible Surface Area (MMPBSA) Method

The MMPBSA method is an approximate method used for determining the absolute free energy difference of binding, $\Delta G_b(s)$, in solution, and makes use of the solvation thermodynamic cycle (see Figure 2.1(b)). As it is used in the quantitative analysis presented in this thesis, specifically in Chapters 6 and 7, we will discuss its basis in some detail here.

As mentioned before, $\Delta G_b(s)$ is calculated indirectly via Equation 2.50. The MMPBSA method implements this by dividing the calculation of the free energy into three main components:

$$\Delta G_b(s) = \Delta G_b^{MMPBSA} = \underbrace{\Delta G^{MM}}_{\Delta Gb(v)} + \underbrace{\Delta G_{pol}^{sol} + \Delta G_{nonpol}^{sol}}_{\Delta G^{sol}} \tag{2.54}$$

The ΔG^{MM} term represents the free energy of binding *in vacuo* ($\Delta G_b(v)$), whereas ΔG^{sol} is the solvation free energy difference upon binding and represents the sum of all the other legs in the thermodynamic cycle required to complete the cycle. It is composed of a polar ΔG_{pol}^{sol} and a non-polar term ΔG_{nonpol}^{sol}. Each component of the free energy is calculated by post-processing the coordinate trajectory produced from a molecular dynamics simulation.

The components of the free energy are calculated at each desired timestep across the trajectory and then averaged over the set of timesteps used. Each of the components in Equation 2.54 is calculated via a different method. Furthermore the components themselves are all constructed from the difference in energies between the complex and the sum of the protein and ligand separately:

$$\Delta G^{MM} = G_{com}^{MM} - (G_{pro}^{MM} + G_{lig}^{MM}) \tag{2.55}$$

for the molecular mechanics components and:

$$\Delta G^{sol} = \Delta G_{com}^{sol} - \Delta(G_{pro}^{sol} + \Delta G_{lig}^{sol}) \tag{2.56}$$

for the solvation components of the free energy.

There are two principal ways of performing the molecular simulation. Either a single molecular dynamics simulation is conducted, from which the trajectories of the complex, the protein and the ligand are extracted for post-processing, or three separate simulations are conducted, one for the complex, protein and ligand, thus naturally generating three separate trajectories. There are advantages and disadvantages associated with either approach. A significant part of a molecular system may not contribute to the free energy difference of binding. Ideally one would want to only include the contribution to binding from those parts of a molecule that induced a change in the free energy. Using a single simulation approach, as the same coordinates are used for the complex as well as for the protein and ligand separately, there is an exact cancellation of terms from those parts of the system which do not influence binding. This cancellation of errors does not occur in the three-simulation approach as each system is free to explore different trajectories and will in general induce a contribution to the free energy difference from each atom. However, using the single simulation approach does not allow as thorough sampling of the conformational differences associated with each species. For example the ligand's conformational behaviour may differ significantly when unbound as compared to bound and these conformations may not be readily accessible in a simulation carried out only in the bound state.

The MMPBSA method is much less computationally intensive than TI or FEP. However, although the method requires no experimental fitting, it is an approximation composed of several assumptions, each of which that lead to terms that make a contribution to the overall free energy of binding.

We will now address how each of the components of the free energy are calculated. ΔG^{MM} is determined by the sum of the electrostatic, the van der Waals and the internal molecular mechanics interactions between the atoms of the protein and the ligand:

$$\Delta G^{MM} = \Delta G^{MM}_{ele} + \Delta G^{MM}_{vdW} + \Delta G^{MM}_{int} \tag{2.57}$$

All of these contributions are attained simply by calculating the interatomic interaction energies for the complex, protein and ligand and making use of Equation 2.55. In the single trajectory method the internal molecular mechanics energy is trivially zero as there is an exact cancellation of terms between the complex, protein and ligand.

Calculation of ΔG^{sol}, however, is more involved. The free energy of solvation of a solute is the change in the free energy of a system upon it being taken from a vacuum into the solvent. This change in free energy comprises both changes in the polar and non-polar interactions between the solvent and the solute.

The polar component, ΔG^{sol}_{pol}, of the solvation free energy is calculated by treating the solvent implicitly as a medium of high dielectric constant and the solute explicitly as a region of low dielectric constant. As such, the system can be treated by the Poisson equation, which relates the electrostatic potential at a point $\phi(\mathbf{r})$ to the charge density $\rho(\mathbf{r})$. Incorporating the effects of ions present within the solvent medium requires an additional component, described by a Boltzmann distribution. This leads to a combined Poisson-Boltzmann description resulting in the Poisson-Boltzmann equation:

$$\nabla.\varepsilon(\mathbf{r})\nabla\phi(\mathbf{r}) - \kappa' sinh[\phi(\mathbf{r})] = -4\pi\rho(\mathbf{r}) \tag{2.58}$$

where, $\varepsilon(\mathbf{r})$ is the dielectric constant which varies from solute ($\varepsilon \sim 1-4$) to solvent ($\varepsilon \sim 80$ for water) and κ' is a constant which represents the ionic strength. By expanding the hyperbolic sine function as a Taylor series and taking only the first term, we arrive at the linearised Poisson-Boltzmann equation:

$$\nabla.\varepsilon(\mathbf{r})\nabla\phi(\mathbf{r}) - \kappa'\phi(\mathbf{r}) = -4\pi\rho(\mathbf{r}) \tag{2.59}$$

Equation 2.59 is solved, firstly by superimposing a cubic grid onto the system such that the values of the electrostatic potential, dielectric constant and ionic strength are assigned at each grid point. The grid points which lie within the solute are then assigned the dielectric constant of the solute, whilst the atomic charges of the solute are distributed amongst neighbouring grid points weighted by the proximity to each neighbour. Finite difference methods are then used to determine the derivatives in Equation 2.58 and, as the potential at each grid point influences the potential at each neighbouring grid point, the

calculation is iterated until the potential converges below a specified tolerance. The polar component, ΔG^{sol}_{pol}, of the solvation free energy is then given by:

$$\Delta G^{sol}_{pol} = \frac{1}{2}\sum_i q_i(\phi_i^{80} - \phi_i^1) \tag{2.60}$$

where q_i is the charge at each point i on the grid and where ϕ is calculated twice, once in which the dielectric constant ε for the solvent is set to 80 to represent water and the other in which ε is set to 1 to represent the vacuum. This calculation is done for each of the complex, protein and ligand species, and by making use of Equation 2.56, the total ΔG^{sol}_{pol} is determined.

The non-polar component, ΔG^{sol}_{nonpol}, of the solvation free energy consists of both the van der Waals interaction, ΔG^{vdW}_{nonpol}, between the solvent and the solute as well as an extra term, ΔG^{cav}_{nonpol}, due to the free energy change of forming a cavity in the solvent. When a solute is immersed in solvent, the required cavity formed in the solvent to accommodate the solute requires work to be done by the solute against the solvent pressure. Additionally, there is an entropic penalty associated with the reorganisation of solvent molecules around the solute.

The solvent molecules in the first solvation shell are the most affected by the reorganisation. As the number of solvent molecules in this first shell is approximately proportional to the solvent accessible surface area (A) around the solute, and as the van der Waals interaction falls off rapidly so that it is also approximately proportional to the number of solvent molecules in the first shell, the non-polar component of the interaction can be modelled linearly, such that:

$$\Delta G^{sol}_{nonpol} = \Delta G^{vdW}_{nonpol} + \Delta G^{cav}_{nonpol} = \gamma A + b \tag{2.61}$$

where the terms γ and b are empirically determined [67, 68]. ΔG^{SA} is then calculated for the complex, protein and ligand and the difference again obtained between the complex and the sum of the separate protein and ligand contributions.

The Configurational Entropy Contribution

Although the MMPBSA method includes the entropic changes upon solvation through the use of the cavity term in the non-polar solvation component, it does not take into account the change in the configurational entropy associated with ligand binding in the gas phase. In general, when a ligand binds to a protein, the increased conformational restriction upon binding of both the ligand and the protein reduces the configurational entropy of the complex as compared to the sum of the unliganded species. The decrease in the configurational entropy is thus a free energetic barrier to binding. The effect can be quite significant and it is thus important to include it in an assessment of the absolute free energy difference of binding.

The configurational entropy is composed of three components, translational, rotational and vibrational, which are summed to give the overall contribution:

$$S^{conf} = S^{conf}_{tra} + S^{conf}_{rot} + S^{conf}_{vib} \tag{2.62}$$

There are three rotational and three translational degrees of freedom for each molecular species, which are restricted upon formation of the complex. Additionally there are 3N-6 degrees of vibrational freedom, where N denotes the number of particles in the system. The free energies of the translational, rotational and vibrational components of configurational entropy are well described from the principles of statistical mechanics [69, 70] and given as follows:

$$S^{conf}_{tra} = \frac{3}{2}RT - RT\left[\frac{5}{2} + \frac{3}{2}ln\left(\frac{2\pi m k_b T}{h^2}\right) - ln(\rho)\right] \tag{2.63}$$

$$S^{conf}_{rot} = \frac{3}{2}RT - RT\left[\frac{3}{2} + \frac{1}{2}ln(\pi I_A I_B I_C) + \frac{3}{2}ln\left(\frac{8\pi^2 k_b T}{h^2}\right) - ln(\sigma)\right] \tag{2.64}$$

$$S^{conf}_{vib} = \sum_{i=1}^{3N-6}\left[\frac{1}{2}h\nu_i + \frac{h\nu_i}{e^{h\nu_i/k_bT}}\right] - \sum_{i=1}^{3N-6}\left[\frac{h\nu_i}{e^{h\nu_i/k_bT}} - RTln(1-e^{-h\nu_i/k_bT})\right] \tag{2.65}$$

S^{conf}_{tra} and S^{conf}_{rot} thus depend on entirely known quantities, where ρ is the number density at 1 mol/L, m, the total mass, I_A, I_B and I_C are the principal moments of inertia and σ, the symmetry factor of the molecule. S^{conf}_{vib} however, depends on the normal modes ν_i of the molecule.

Normal mode analysis is therefore one of the principal methods used to determine the loss of configurational entropy upon binding [70–72]. The normal modes of a macromolecule are a set of concerted motions of its constituent atoms that behave harmonically close to a minimum energy point. The lower the frequency of the oscillation, the larger the amplitude of the normal mode. Normal modes are therefore useful in separating slow macroscopic vibrations of a protein from the rapid vibrations of say a hydrogen atom. However, one of the fundamental limitations of normal mode analysis is the assumption that the oscillations of the molecule occur within a single potential energy well. In fact, the energy landscape of a macromolecule is very rugged, which severely limits the conformational space over which normal mode analysis can be performed. Conformational changes in a protein may thus render normal mode analysis ineffective. Attempts have been made to address this through the development of alternative methods such as quasi-harmonic analysis [72, 73], which allow for sampling over several neighbouring potential wells.

Calculation of normal modes is computationally demanding due to the requirement to determine and then diagonalise the Hessian matrix, a matrix containing the partial second derivatives of the potential with respect to atomic positions. This calculation scales as N^2, and is thus the limiting factor of the rate at which the calculation can be performed. For a thorough treatment regarding the theoretical or computational basis of calculating normal modes the reader is referred to the literature [74].

Once obtained, the normal modes of a molecular species can be converted into the vibrational component of the configurational entropy as described in Equation 2.65. The configurational entropy is then calculated for the complex as well as the protein and ligand separately and again the difference between the complex and the sum of the protein and ligand values determines the change in the entropy:

$$\Delta S^{conf} = S^{conf}_{com} - (S^{conf}_{pro} + S^{conf}_{lig}) \tag{2.66}$$

The entropic contribution can in turn be subtracted from the free energy obtained from an MMPBSA calculation to determine a more accurate absolute free energy of binding:

$$\Delta G_b = \Delta G^{MMPBSA}_b - T\Delta S^{conf} \tag{2.67}$$

The NMODE module in the AMBER 9 package [53] contains an implementation of the calculation of the entropic contribution using normal mode analysis. In Chapters 6 and 7 we apply both the MMPBSA and the normal mode methods in order to determine absolute free energies of binding of ligands to HIV-1 protease.

2.6 High Performance Computing (HPC) and Grid Technology

As discussed previously, the timestep used in MD simulations is typically around 1 fs and the requirement to model systems accurately means that simulating proteins in explicit solvent is often necessary. Such requirements greatly increase the number of particles that need to be simulated and thus makes molecular dynamics methods very computationally intensive. For example, using the molecular dynamics package NAMD, it would currently take about 250 hours to compute 1 ns of a 30,000 atom system (which describes a small protein of around 200 amino acids in explicit solvent) on a single 1.3 GHz Itanium processor.

It is clear that this high computational demand is a great limiting factor on serial processors. The ability therefore of utilising high performance computing resources coupled with codes that parallelise the computation of large scale systems is of enormous benefit to this field (see § 2.3.4). This has led to the simulation of larger systems for a longer period of time with the 'state of the art' currently approaching 1 million atom systems of biomolecules bound to lipid membranes in explicit solvent over timescales of tens of nanoseconds [75].

Efficient scaling allows proportional speed up with the use of more parallel processors, but is eventually limited by the increased communicational loads produced from excessive numbers of processors. There exists, typically an optimum number of processors to use given a system size, before departure from linearity is observed. The ability to scale well is therefore a crucial feature of any MD code and in

this capacity, the performance of packages such as CHARMM and AMBER is limited [26, 33], whereas LAMMPS and NAMD offer far superior scaling capabilities [34, 36]. For example, 1 ns of temporal dynamics of the same 30,000 atom system can be computed on 32 parallelised 1.3GHz Itanium processors in under 8 hours using NAMD as the molecular dynamics package. Implementations of molecular dynamics codes are currently available on many HPC resources. Examples of HPC resources are the HPCx machine at Daresbury in the UK [1], providing 2560 1.5 Ghz IBM Power5 processors and the Lonestar machine at the Texas Advanced Compute Centre (TACC) in the USA [2], providing 5200 2.6 Ghz Xeon 5100 series processors, amongst many others.

However, whilst many HPC resources exist and can be utilised independently, an even greater computational benefit can be afforded through a combined availability and use of several resources simultaneously. Furthermore, there are often several overheads associated with the effective implementation of simulation 'jobs'. These include long queueing times before a job begins to run as well as required knowledge of resource dependent job submission scripts.

It is in this context that the concept of a computational 'grid' is perhaps the most appealing [76]. The definition of a computational grid, analogous to an electricity grid, as a "network of high performance computational resources accessible through middleware that allows for the seamless submission and execution of simulations across administrative domains" [77], provides excellent potential benefit to high performance computing in general, as well as its application to molecular simulation . There are currently several grids in operation around the world. Amongst these are the National Grid Service (NGS) in the UK [3], the US TeraGrid [4] and the Distributed European Infrastructure for Supercomputing Applications (DEISA) Grid [5] each of which combine the HPC resources of several sites within their respective domains.

2.6.1 The Application Hosting Environment

From the perspective of a user, the added benefit of using a grid, as opposed to independently accessing several HPC resources, is severely dependent on the seamlessness of the middleware employed to administrate the submission and execution of computation around its constituent component resources. In reality, the use of grid services in the manner described above is still far from seamless, largely due to the overriding complexities of the middleware designed to facilitate such grid usage.

Several implementations of middleware exist for grid computing, such as the Globus Toolkit [78] and Unicore [79]. However, it is becoming increasingly apparent that the use of such middleware is too complex for general scientific purposes and this is severely hampering the uptake of grid computing

[1] http://www.hpcx.ac.uk
[2] http://www.tacc.utexas.edu/resource/hpcsystems
[3] http://www.ngs.ac.uk
[4] http://www.teragrid.org
[5] http://www.deisa.org

[80].

In particular, conventional grid middleware is characterised as being 'heavyweight', due to the significant obstacles in the way of the scientist, prior to successful deployment and use. This is especially the case for the grid-middleware's client side software, which a user has to interface with when attempting to utilise grid services. The heavyweight nature of grid middleware has motivated the recent development of several 'lightweight' middleware solutions [81, 82] that attempt to reduce the complexity of utilising grid resources.

One such recently developed solution is the Application Hosting Environment (AHE) [83]. The AHE is a lightweight environment for hosting scientific applications, and works by exposing such applications as web services. It is novel in the sense that it is application-centric as opposed to job-centric and allows users to run unmodified applications on grid resources where each job creates instances of the application, as well as managing file transfers and job submissions and allowing job monitoring. Whilst use of the AHE requires an appropriate grid middleware on the grid resource to be installed, such as Globus or Unicore, no such requirement is necessary on the client side. This greatly facilitates the use of grid resources for the scientist without the overhead of having to deal with complex middleware or of having to 'shepherd' simulations manually from resource to resource. Applications currently hosted by the AHE include molecular dynamics codes such as NAMD2 [34], DL-POLY [35], LAMMPS [36] and the lattice-Boltzmann code, LB3D [84].

An additional benefit of the AHE is that alongside the graphical user interface (GUI), which facilitates submission of jobs, a command line interface also provides the same facilities. The latter can easily be executed from Perl scripts, through which workflows can be constructed for the implementation of many simulations in a particular order. Furthermore due to the single uniform interface with resources in multiple administrative domains, the AHE can be used to marshall workflows across many HPC resources in a true grid sense. Workflows are discussed further in § 2.6.2. An example of how the AHE has been utilised in this manner is provided in Appendix A, where we describe a tool, called the 'Binding Affinity Calculator' (BAC), designed for the automated calculation of binding free energies of HIV-1 protease-ligand variants.

2.6.2 Exploiting High Performance Computing and Grids

The ever increasing supply of high computational power as well as the technological advances afforded by grid computing, albeit with the use of suitable middleware, allows for novel approaches to be designed and adopted for the enhanced conduction of computational simulations.

Enhanced Computational Steering

Computational steering is an attractive prospect that can have enormous benefit (see § 2.5.1). The ability to intervene in a simulation by 'steering' it in a desired direction is clearly a huge advantage over being dependent on the final outcome of a simulation, before assessing whether it needs to be repeated in order to better study phenomena of interest that may emerge through the course of the simulation. This can be facilitated by saving the state of a simulation or 'checkpointing' regularly, such that it can be restarted from the desired checkpointed state, possibly with different parameters to explore the outcomes of the simulation further.

The benefit of grid computing is enhanced when combined with computational steering. An example of this has been under the Simulated Pore Interactive Computing Environment (SPICE) project [85]. The translocation of DNA, RNA and polypeptides across trans-membrane protein channels such as α-hemolysin are important biological events to understand. Unfortunately, such events occur on the μs timescale and are intractable using conventional molecular dynamics.

The use of steered molecular dynamics approaches combined with the use of Jarzynski's equation [86, 87], allows the equilibrium free energy profile of the translocation to be calculated in the presence of non-equilibrium forces. The DNA molecule is steered along the vertical channel axis by a 'dummy' atom connected to the DNA which can vary in its steered velocity v, as well as the strength of its connective spring constant k (see § 2.5.1). Steering these parameters allows for the optimal free energy profile to be calculated. Furthermore, the use of grid resources allows the single long simulation to be partitioned along the translocation axis into multiple shorter simulations, each of which can be distributed and run in parallel across various grid resources. Consequently, the μs timescale can be explored, whilst only using the computational time required to simulate the ns timescale in conventional equilibrium molecular dynamics simulations.

Improving the Turnover of Free Energy Calculations

Another example of the enhanced benefit afforded by coupling grid technology to computational steering has been in the massive speed up of binding affinity calculations that use TI. Conventionally the two sets of TI simulation runs needed for a $\Delta\Delta G$ calculation are performed sequentially (see § 2.5.2). Computational steering on a grid offers the ability to run almost simultaneous simulations that are started by the prerequisite of equilibrating the system before initiating the production run from which ensemble averages are calculated. As a consequence, turn-over time of a calculation is rapidly increased with a complete calculation being feasible in days rather than weeks or months. This method, known as 'Steered Thermodynamic Integration using Molecular Dynamics' (STIMD), has been applied to the study of the inhibition of SH2 domains, protein domains involved in the complex signalling pathways of many processes in biological organisms [61].

Workflow Management

The management of workflows offers extended benefit to computational simulation through the construction of simulation components that follow a particular order or sequence. Whilst computational steering allows interactivity with the simulations, workflow management is more geared towards an automated approach to conducting a set of simulations, where the course which a simulation adopts is based on the outcome of the previous step within the workflow.

For example, the molecular dynamics simulations implemented in Chapters 4-7 all consist of several component simulations conducted sequentially, with the output of one serving as the input of the next in the chain.

Beneficially, this linear example of workflow as well as others are readily implementable within the infrastructure afforded by the AHE, discussed above. The user does not need to manage each of the simulations, as they are automatically conducted and transferred from grid resource to a pre-specified location. In Chapter 6 and Appendix A we will discuss how workflows have been implemented using the AHE, in order to develop an infrastructure for conducting automated free energy calculations of drugs and substrates binding to HIV-1 protease.

2.7 Alternative Computational Methods for Studying Proteins

We now outline the main alternative computational methods that have been developed for protein structure determination as well as for evaluating the strength of protein-ligand binding. The examples that will be discussed here are homology modelling, *Monte Carlo* (MC) methods and *ab initio* methods. All employ different methodologies and are applied to different problems; homology modelling is used for structure determination and is based on the knowledge of large pre-existing data sets of known structure. MC is a stochastic method that is best used for the generation of large ensembles of thermodynamic data allowing calculation of binding affinities, whilst *ab initio* methods are more directly concerned with electronic structure calculations.

2.7.1 Homology Modelling

Homology modelling is a method that is applied to the protein structure prediction problem. It is an inductive bio-informatics approach that infers the structure of proteins from homologues of known structure. Homologues are proteins that have amino-acid sequence identities in a significant number of positions along their polypeptide chains, such that they are considered to have evolved from a single common ancestor. The level of identity required to be considered a homologue is determined through statistical methods that make comparisons with computer-generated random sequences [5].

Often, two proteins that were below the threshold of statistical significance to be considered ho-

mologues, have later been determined to have similar structures and functions. The requirement of statistical significance is then only a good criterion for homology if two similar sequences don't yield very different structures. The crucial pre-requisite of homology modelling is the existence of homologues. There exists a vast divergence between the number of determined sequences and structures of proteins, with far more sequences having been determined than structures. It is often the case that suitable homologues cannot be found.

If suitable homologues do exist, the target protein is aligned to the appropriate homologue template from which secondary structure can usually be mapped and then validated. Following this, a model can be built that estimates the tertiary protein structure. This employs some form of energy minimisation subject to restraints. There are a plethora of homology modelling resources available catering for different aspects of the modelling process, both as web accessible servers such as GenBank[6], which serves as a repository of different sequences and as stand-alone programs such as MODELLER[7], that build the tertiary structure following alignment. We will not list all of them here; for a comprehensive account the reader is referred to the literature [88].

Homology modelling is a very important tool for the determination of protein structures. It is, however, limited by the existence of structurally known homologues and these are usually determined by X-ray crystallography, under unphysiological conditions. The errors in homology modelling that occur because of this can be reduced if the method is combined with MD, which can take a starting protein structure and evolve its dynamics to explore further equilibrium structures. This can then aid the refinement of the homology data set to yield more accurate results.

2.7.2 Monte Carlo Methods

The Monte Carlo simulation method was historically the method used in the first simulation of a molecular system [22]. Unlike molecular dynamics however, MC does not follow the time evolution of a system and is thus not deterministic. It uses a stochastic approach where the configuration of a system is generated by making random changes to the positions of the particles present. A special set of criteria then decide whether the new configuration is accepted or not.

The description of systems in a thermodynamically correct way requires the use of the Metropolis MC method [89]. In such a method, acceptance criteria ensure that the probability of obtaining a configuration is Boltzmann-weighted and leads to the generation of what is called a Markov chain of states. These satisfy the condition that the outcome of each trial is dependent upon only the previous trial and that each trial belongs to a finite set of possible outcomes.

Monte Carlo methods are therefore used in energy minimisation problems as well as being used to calculate the thermodynamic properties of systems through the generation of large ensembles. This

[6] http://www.ncbi.nlm.nih.gov/GenBank

[7] http://guitar.rockefeller.edu/modeller/modeller.html

makes them particularly suited for calculating binding affinities between enzymes and inhibitors and for the refinement of structural predictions from homology models.

2.7.3 *Ab Initio* Methods

Ab initio refers to calculation from the first principles of quantum mechanics. *Ab initio* methods are thus the attempt to model many-body systems by determining approximate solutions to the time dependent Schrödinger equation:

$$\left(-\frac{\hbar^2}{2m}\nabla^2 + V\right)\Psi(\mathbf{r},t) = i\hbar\frac{\partial\Psi(\mathbf{r},t)}{\partial t} \tag{2.68}$$

where $\hbar$ is Planck's constant divided by 2π, m is the mass of a particle and $\Psi(\mathbf{r},t)$ is the wavefunction of the system. We will not go into any detail regarding the solution of this equation here. For the purpose of this thesis it suffices to say that solving the Schrödinger equation for a system of particles described by the function $\Psi(\mathbf{r},t)$ allows for the time evolution of the quantum system. However, the equation is analytically unsolvable for any system beyond a Helium ion (He^+).

In practice there are many approximations that are made to facilitate solving the Schrödinger equation for multi-atomic and multi-electronic systems, such as the Born-Oppenheimer approximation which separates the contribution of the nuclei from the total wavefunction. In these methods the groundstate of the electronic wavefunction is calculated to give the forces on the nuclei. These forces then evolve the positions of the nuclei of the atoms using classical mechanics, assuming that the electrons will remain in a ground state, following the nuclear movements instantaneously. The wavefunctions are then recalculated to determine the new ground state and so on.

There are a whole host of quantum mechanics based methods for calculating the wavefunction that vary from being very computationally intensive, to more empirical methods such as Density Functional Theory (DFT), or Hartree-Fock basis set expansions. Several of these can be employed by QM software such as Gaussian '98 [90].

Quantum mechanical methods are the most computationally demanding methods in computational science and it is unfeasible to model protein dynamics using them. As this thesis focusses on the dynamics of macromolecular systems, intractable using the QM methods alluded to above, we will not look in any detail into quantum mechanical methods here. For a comprehensive treatment of the subject, the reader is referred to the literature [22].

However, even though quantum dynamics is not easily tractable, it is easier to perform energy minimisation on smaller structures such as inhibitors or ligands and to calculate optimum geometries. It is also possible to fit electrostatic potentials to point charges for novel molecules whose classical force-fields are undefined and these can subsequently be used in classical molecular dynamics applications. An implementation of this has been carried out for the studies reported in Chapters 4, 5 and 6.

It is also possible to use QM based methods to explore the chemical nature of enzymatic processes [91]. This has been optimally achieved using methods that combine QM and classical molecular dynamics methods (QM/MM) and is discussed further in § 2.8.1.

2.8 The Limits of Molecular Dynamics

Molecular dynamics is an implementation of the description of polyatomic interactions that partially takes into account the chemical nature of molecular, organic and biomolecular structures by representing them as components (bonds, angles, dihedrals etc.) in a classical analytical potential. As mentioned before, the parameters for this potential are empirically derived for small systems and transferred to a set of larger structures. This methodology has worked and continues to work rather well for a whole host of problems that involve the equilibrium and sometimes non-equilibrium dynamics of systems as well as calculations of free energy and other thermodynamic properties.

MD is limited, however, in its representation of matter, by being too course grained to study the fundamental chemical/physical nature of certain processes, such as chemical reactions, that involve an understanding of the structure and dynamics of electrons. Conversely it is too atomistically detailed to study processes that occur on larger scales such as hydrodynamics. For example, MD cannot properly describe bond formation or bond cleavage, yet the function of many biomolecular systems, such as enzymes, is just this. Such processes are the realm of *ab initio* quantum mechanical methods (see § 2.7.3).

On the other end of the spectrum, the computational requirement of implementing MD is such that the cutting edge is currently $10^5 - 10^6$ atoms on large parallel supercomputers with achievable timescales of tens of nanoseconds. It is therefore completely unfeasible to use MD to simulate the flow of a river, for example. This has typically been the realm of computational fluid dynamics. It is also unfeasible to study many higher order biological processes such as the internal mechanisms of a cell or even the full workings of a membrane including its host of membrane proteins as the spatial and temporal scales that these processes exist on is currently too large to be handled by current computational power.

2.8.1 Multi-Scale Modelling

What emerges from these limitations on molecular dynamics is the idea of integrating MD with other methods of description to achieve 'hybrid' descriptions of a system that traverse multiple spatial and temporal scales and, to this end, coupled models have been developed for MD.

QM/MM

Molecular dynamics does not afford an electronic description of matter and is this unable to model chemical reactions. Whilst an electronic description of matter can be dealt with using *ab initio* quantum

mechanical methods that solve the Schrödinger equation for small molecules (see § 2.7.3), such methods are intractable to the application of biomolecular modelling. However, many biomolecules such as enzymes, are involved in chemical processes, catalysing a large array of bond formation and cleavage reactions. A tractable description of these processes, for example the formation of transition states in an enzymatic reaction, is thus important in further understanding protein function.

QM/MM methods couple *ab initio* methods with classical molecular dynamics. The region of interest in the biomolecule is then described by the more computationally demanding QM method, whilst the majority of the system, including water molecules are described by classical MD. A description of the interaction between the two regions is additionally necessary to correctly couple the two regions and thus describe the whole system. The Hamiltonian of the system is given by:

$$H_{TOT} = H_{QM} + H_{MM} + H_{QM/MM} \tag{2.69}$$

Whilst such a method is attractive, due to the obvious computational enhancement over using just *ab initio* methods, accuracy of representation is dependent on the schemes used to couple the two regions. For reactions in which reactant species can be completely decoupled from the classical region, this is relatively easy. For example, the conversion of chorismate to prephenate, catalysed by chorismate mutase does not involve covalent bond formation or bond breaking with the enzyme. Chorismate can thus be completely treated quantum mechanically, whilst the enzyme as a whole can be treated classically [92].

Reactions in which the macromolecule is explicitly involved in forming and breaking covalent bonds are more difficult to model as they require the partitioning of the two regions across bonds. An example is the suggested mechanism for the peptide hydrolysis of substrates cleaved by HIV protease (see Chapter 3), which involve alteration of the protonation states of the catalytic aspartic acids of the enzyme as well as transitional coupling to the substrates. As the treatment of the interface between these two regions is highly non-trivial, there is still much room for development and the reader is referred to the literature for an account of various strategies that are employed [91, 93, 94].

Hybrid MD

Hybrid MD is a novel method that couples classical molecular dynamics with a continuum fluid dynamics region. This is a very powerful method that has the potential to achieve much speed up over classical MD. For example, the accurate simulation of biomolecular systems requires an account of solvation interactions. Water can affect the dynamics of a protein through long range electrostatics, hydrogen bonding and entropic contributions to free energy and is thus incorporated as fully atomistic explicit solvent in MD simulations. However, the drawback is that most atoms in the simulation are then solvent atoms and whilst an atomistic description of the solvent close to the protein is desirable, it

would also be desirable to describe the long range interactions of the solvent in a more implicit and thus computationally efficient way.

Hybrid MD addresses this by coupling an MD region to a surrounding hydrodynamic region conserving the correct transfer of energy, momentum, pressure and number density across the interface. Particles leaving the MD region are incorporated into the fluid description of the continuum region and based on the properties of the continuum region, particles are also inserted into the MD region using an algorithm known as USHER. For a more thorough account the reader is referred to the literature [95–97].

Grand Canonical Molecular Dynamics (GCMD)

As mentioned before, it is possible to set up an MD simulation in a number of different thermodynamic ensembles, for example, the microcanonical (NVE) and canonical (NVT) ensembles. However, many biological systems operate in a way which does not preserve particle number at equilibrium, but rather the chemical potential μ of the system. Thus an ideal capability for an MD simulation would be the correct set up of what is known as the grand canonical ensemble (μVT).

Conventionally the problem with this has been the difficulty of inserting or removing particles from the MD simulation. Inroads have been made in this regard. Lynch and Pettitt [98] described a semi-grand canonical MD in which only the solvent atoms varied and used this to describe the bovine pancreatic trypsin inhibitor (BPTI). The recent development of algorithms such as USHER coupled with its extension to the insertion of polar solvent such as water [96, 97], may facilitate the development of a full GCMD that may eventually be applied to a whole host of biological systems.

CHAPTER 3

The Human Immunodeficiency Virus

VIRUSES are one of the simplest and most effective examples of life. As parasitic organisms, they utilise the translational machinery of cellular organisms to produce multiple copies of themselves and thus propagate. In this chapter, we aim to provide an overview of the human immunodeficiency virus (HIV) by outlining various aspects of its emergence, its biological structure and life-cycle and its course within infected individuals. We will also comment on the treatments that have been developed to combat its progression. As this thesis focusses on one of the key enzymes of HIV, the HIV-1 protease, we subsequently discuss the structure and function of this enzyme as well as the extensive experimental and computational studies that have been undertaken to elucidate its structural, dynamical and functional properties. We finally discuss some of the inhibitors that have been developed to target the protease as well as outlining the emergent problem of drug resistance which forms the key subject matter of later chapters in this thesis.

3.1 The Emergence of HIV

The acquired immunodeficiency syndrome (AIDS) was first diagnosed in the United States in 1981 following unusually high occurrences of rare diseases such as Kaposi's sarcoma and several diseases associated with immunocompromisation such as *Pneumocystis pneumonia* and *lymphadenopathy* (non-Hodgkins lymphoma) [99, 100]. It was obvious from the spread of the disease (primarily through blood transfusion or sexual intercourse) that some infectious agent was the cause, and a selective loss of T helper cells suggested a virus, but it was not until 1983 that HIV was discovered to be the causative agent of AIDS. This came about with the ability to grow the virus *in vitro* [101].

HIV is a retrovirus. These are a class of virus that contain genetic information in the form of RNA which encodes for the enzyme reverse transcriptase (RT). RT converts viral RNA into DNA which is then integrated into that of the host cell's. Within this class of viruses, HIV is classified as a lentivirus and has genetic and morphological similarities to animal lentiviruses such as simian immunodeficiency virus (SIV), feline immunodeficiency virus (FIV) and bovine immunodeficiency virus (BIV) [102, 103]. What

separates lentiviruses from other retroviruses is the complexity of their viral genomes. In addition to the three genes *gag*, *pol* and *env* contained by standard retroviruses, which encode matrix/capsid proteins, enzymatic proteins and envelope proteins respectively, lentiviruses contain additional regulatory genes that probably contribute to the increased pathogenicity which differentiates them from other retroviruses [104]. Lentiviruses often cause immunodeficiency in addition to slow, progressive wasting disorders, neurodegeneration and death [105]. Viral protein cross-reactivity and sequence similarities have shown that HIV is especially close to SIV [106]. It is not clear exactly when HIV entered the human population, but it is thought that a chimpanzee species of SIV (SIV_{cpz}) might have jumped to the human population in the Congo. The first known human infection of HIV was that retrospectively found in a blood sample taken from a man in the Congo in 1959 [101].

HIV is itself classified into HIV-1 and HIV-2. The virus as described above is what came to be classified as HIV-1. It is HIV-1 that is more pathogenic and responsible for the worldwide global epidemic [101]. HIV-2 is similar to HIV-1 but shares greater genetic similarity with SIV. Like SIV, HIV-2 contains an extra gene that encodes for the viral protein *vpx*, which is not contained in HIV-1. It is thought that HIV-2 entered the human population in the 1940's and probably came from a different primate species, possibly the sooty mangabey monkey which exists on the West African coast. This region is the centre of the more localised HIV-2 epidemic [107]. HIV-1 is further subdivided into several subtypes, within two groups M (Main) and O (outlier). Within Group M, the subtypes are classified as A-J. Figure 3.1 shows the distribution of the main subtypes around the world.

In 2006, approximately 39.5 million people were estimated to be living with AIDS, with approximately 4.3 million being newly infected with HIV in 2006 itself. Since 1981, approximately 65 million people (1% of the world's adult population) have become infected with HIV and 25 million people have died of AIDS including around 2.9 million in 2006. Sub-Saharan Africa is one of the worst affected regions, containing 63% of all people in the world currently living with HIV. More detailed statistics of the current global AIDS epidemic can be found at the UNAIDS website[1] as well as the AVERT website[2]. It is worrying that an infection such as HIV continues to have a large and increasingly adverse effect on the health of the global human population. There are multiple reasons why the HIV epidemic has spread so rapidly and remains a major risk to world health, ranging from social, epidemiological, medical and scientific problems that remain to be overcome. This thesis is solely concerned with scientific aspects of HIV treatment and in the following sections we provide an overview of its life-cycle within an infected individual and thus provide a short account of the medical and scientific problems which arise in treating it.

[1] http://www.unaids.org

[2] http://www.avert.org

Figure 3.1: Map showing the distribution of HIV-1 subtypes A-E around the world. Major prevalence of a subtype is denoted in upper case and minor prevalence in lower case. The main subtype prevalent in North America and Europe is subtype B, whereas subtypes A and C are prevalent in Africa and Asia (reproduced from: Microbiology and Immunology Online [101]).

3.2 The Life-Cycle of HIV and AIDS

We will now provide a brief account of the course of HIV infection in individuals which invariably leads to immune failure and AIDS associated fatalities as well as focussing on the life-cycle of the virus, where we discuss the method by which discrete units of the virus, known as virions, infect target cells in humans and utilise such cellular machinery to produce many copies of themselves.

3.2.1 Course of the Infection

The primary target for HIV is the activated CD4 T4 helper lymphocyte, a crucial component of the human immunological system, although HIV can infect other cell types including macrophages, provided they have CD4 receptors on their surface. CD4 helper T-cells bind to B-cells where they are presented with epitopes (antigens digested into fragments). They then release lymphokines, which trigger the mitosis and differentiation of B-cells into plasma cells which subsequently release their B-cell receptors as antibodies. It is this crucial component of the immune system that is compromised by HIV.

Early infection is frequently followed by an acute clinical syndrome appearing 2-6 weeks after infection, the symptoms of which are usually fever, rash and swollen lymph glands [101]. It is associated with a drop in the numbers of circulating CD4 helper T-cells in conjunction with high levels of viral replication in the peripheral blood (more than 10 billion per day). At this stage there can be between

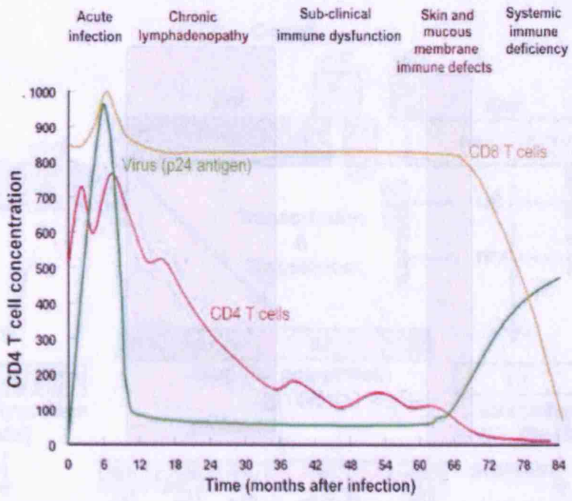


Figure 3.2: Graph showing the evolution of the virus with time as well as the progressive decline of CD4 T helper cells. The course of HIV infection is marked by an initial peak in virions followed by a latency period which can last several years. Eventually the immune system is paralysed and HIV emerges from latency into the onset of AIDS (reproduced from: Microbiology and Immunology Online [101]).

10^4 and 10^7 virions per ml of blood [101].

This is followed by a strong adaptive immune response and initially, a steady state is set up between destruction of infected cells and production of new CD4 helper T-cells. As a result of the immune response the virus largely disappears from the circulation, leading to the partial restoration of CD4 counts in the peripheral blood [101]. It is also at this stage that the virus disseminates to other regions such as the lymphoid and nervous systems.

CD4 helper cells have both an activated state and a resting memory state. HIV replication resulting in the destruction of helper T cells only occurs in the activated state and it is possible for some CD4 T cells to revert back to a resting memory state before destruction. These cells now carry a copy of the HIV genome which will remain latent until the cells are reactivated by antigen. This process is responsible for the next stage of the infection known as clinical latency which can last for several years (see Fig 3.2). During latency, the number of activated CD4 cells gradually declines as they are continually destroyed by either the cytotoxic T-cells of the immune system which ingest the infected helper T-cells or by the virus which destroys helper T-cells. This is through the multiple rupturing of the helper T-cell's membrane achieved by the many virions that bud out of the host cell. The virus itself reaches a constant concentration, the value of which determines the period of latency.

Clinical latency can last up to 15 years, but eventually the immune system fails due to progressive decline of CD4 cells including reactivated CD4 cells containing the latent HIV genome. Cytotoxic T

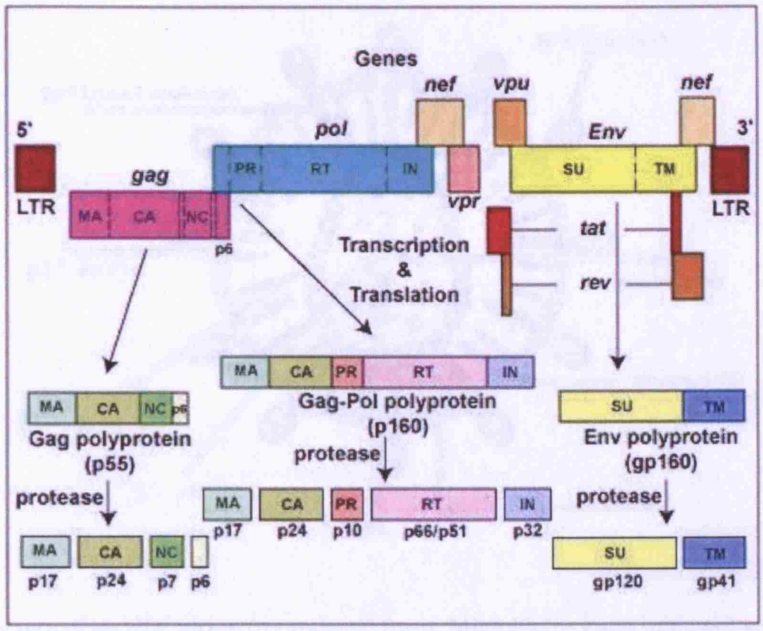


Figure 3.3: The HIV genome
(reproduced from: http://www.cat.cc.md.us/courses/bio141/lecguide/unit2/viruses/hivgenes.html).

cells, normally responsible for the destruction of the virions are not signalled due to lack of helper cells and this results in rapid rise in viral concentration in which the efficacy of the immune system drops to nearly zero. Normally benign opportunistic pathogens are then able to infect the body and it is the emergence of these HIV-associated diseases that characterises the onset of AIDS, which is eventually fatal.

3.2.2 The Structure and Life-Cycle of HIV

As mentioned before, HIV is a lentivirus, a family of retroviruses with a particularly complex genome. The HIV genome (see Figure 3.3) is carried by two identical strands of RNA 9.2 Kb (kilobases) long and contains the standard *gag*, *pol* and *env* genes found in other retroviruses as well as regulatory genes such as *tat*, *rev*, *nef*, *vif*, *vpr* and *vpu*. The *env* gene encodes the envelope associated Env polyprotein gp160 that is later cleaved into two functional proteins gp120 and gp41. The *gag* gene encodes the Gag polyprotein that is later cleaved into four main proteins, *inter alia*, that form the building blocks of the viral core. These are the capsid protein (CA), the matrix protein (MA), the nucleocapsid protein (NC) and protein p6. The *pol* gene is the third standard retroviral gene and combines with the *gag* gene utilising a shift in reading frames, to encode the Gag-Pol polyprotein. This is later cleaved into matrix and capsid proteins as well as the reverse transcriptase (RT), integrase and protease enzymes which are all crucial for the reproduction of HIV.

Figure 3.4 shows the structural arrangement of an HIV virion. Two viral RNA strands are contained

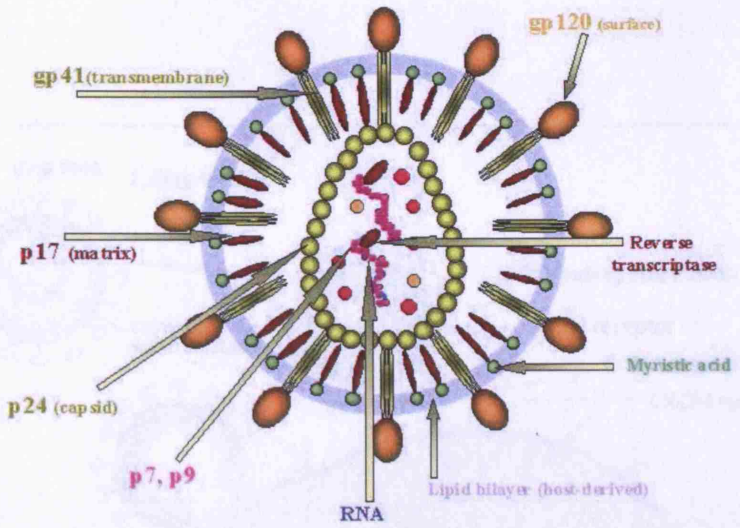


Figure 3.4: Structure of an HIV virion (reproduced from: Microbiology and Immunology Online [101]).

within the nucleus as well as the viral enzymes, reverse transcriptase, integrase and protease. The nuclear shell is made up of capsid proteins (CA). Beyond this lies another shell that is made up of the matrix proteins (MA) and finally, a lipid bilayer membrane encases the virion embedded with gp41 trans-membrane proteins conjugated to gp120 proteins that protrude from the membrane.

Figure 3.5 shows the life-cycle of HIV. HIV infects target cells through the binding of its envelope protein gp120 to the CD4 receptor expressed on helper T cells and macrophages. This binding induces conformational changes that expose certain regions of the gp120 and allow it to interact with a chemokine receptor CCR5 or CXCR4 that is also expressed on the surface of the host cell [108]. This combined interaction in turn triggers a sharp conformational change in the gp41 trans-viral membrane protein which, through a mechanism that is not completely known, brings the viral membrane into proximity with the host cell membrane and thus initiates membrane fusion.

Membrane fusion of viral and host membranes is required for the injection of the virion content into the cell cytoplasm. It is thought that the mechanism by which membrane fusion is initiated is the double backing of gp41 consisting of a coil of 3 alpha helices into a hairpin conformation whilst attached to both membranes [109].

Once injected into the interior of the cell, the viral RNA is reverse transcribed by the reverse transcriptase enzyme (RT) into viral DNA which is then integrated into the host DNA by the viral enzyme integrase [101]. Following this process, a state of latency can occur which may last for the entire lifetime of the cell, unless it receives specific immunological signals by which it is activated. Once activated, the host DNA including the inserted viral DNA is transcribed into RNA. This viral RNA is then transcribed by the standard cell machinery into several viral polyprotein chains, namely the Gag (containing matrix

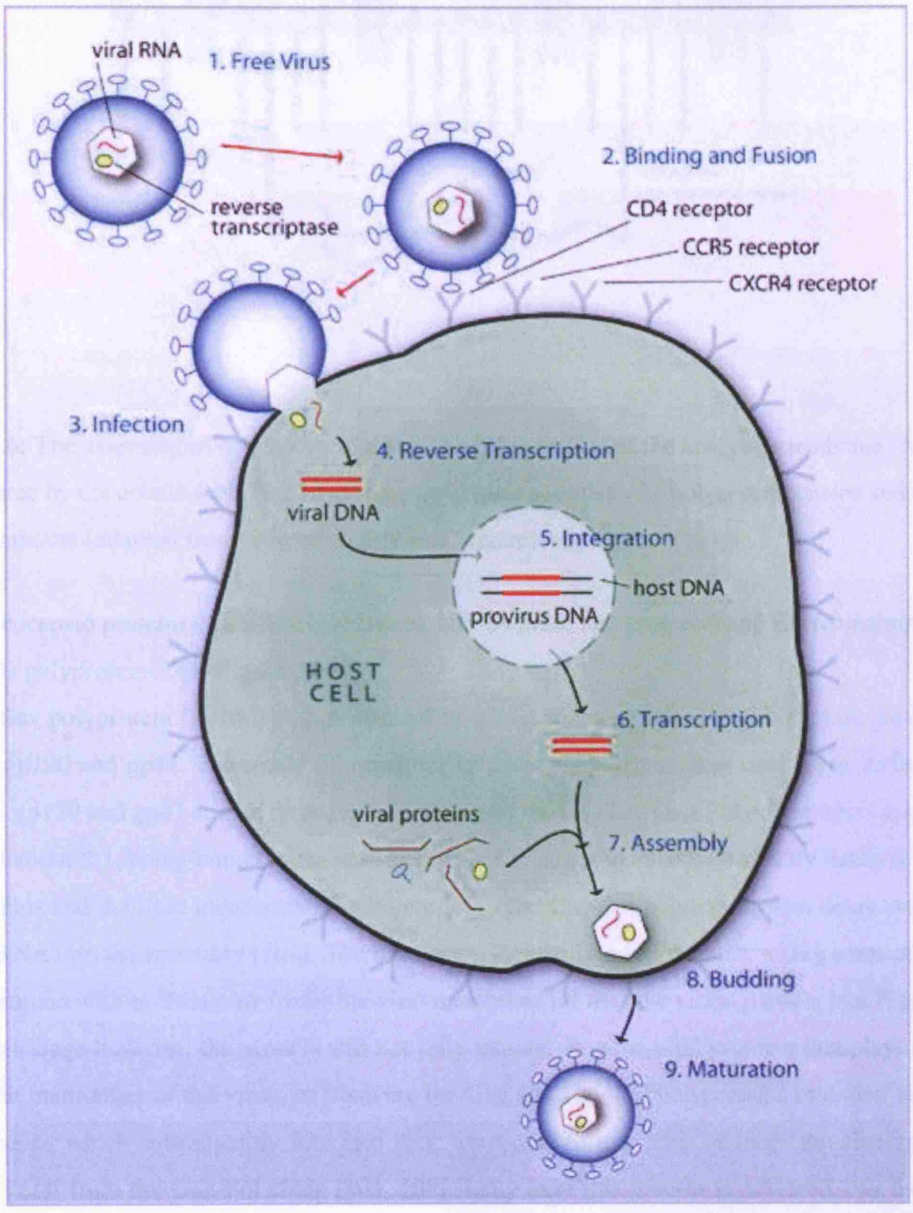
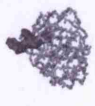


Figure 3.5: The life-cycle of HIV. The main processes are viral entry, reverse transcription into DNA, integration of viral DNA into host, manufacture of viral polyproteins, assembly at the cell surface, budding and maturation through proteolytic cleavage (reproduced from: http://www.dharmacon.com/m360/newsletter/archive/iss2_vol1/images/hiv_med.png).



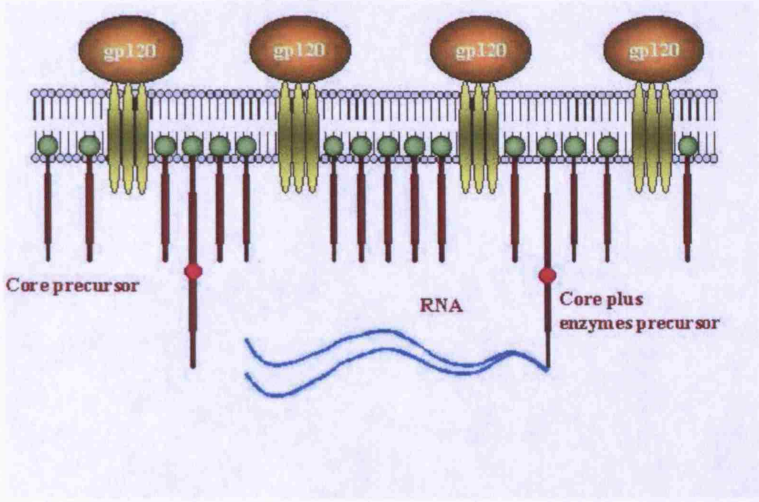


Figure 3.6: The assembly of new polyproteins at the inner surface of the host cell membrane. Assembly is facilitated by the covalent binding of myristic acid (green) both to the polyprotein chains and the inner lipid membrane (adapted from: Microbiology and Immunology Online [101]).

and nucleocapsid proteins), Gag-Pol (containing RT, integrase and protease) and Env(containing gp120 and gp41) polyproteins (see Figure 3.3).

The Env polyprotein (gp160) is then cleaved by a host enzyme in the Golgi body into the envelope proteins gp120 and gp41. Assembly of a new virion takes place at the inner-membrane surface of the host cell. gp120 and gp41 embed through the surface and the Gag and Gag-Pol polyproteins assemble in a ratio of about 8:1, being bound to the membrane by myristic acid, which covalently bonds to both the polyproteins and the lipid membrane (see Figure 3.6). The Gag-Pol polyprotein also drags two strands of viral RNA into the assembly [101]. The new virion then buds out of the host, taking some of the host cell membrane with it. This now forms the viral membrane for the new virion particle (see Figure 3.7).

At this stage however, the virus is still not fully mature. It is the viral protease that plays a crucial role in the maturation of the virus, by cleaving the Gag and Gag-Pol polyproteins into their respective viral proteins which subsequently fold into their functional forms. This includes the cleavage of the protease itself from the Gag-Pol chain [101, 108]. Only once this process is achieved, can the mature virion infect other host cells.

Interestingly, even though this process completes in virions that have already budded out of the host cell, previous studies have shown that initiation of proteolytic activity is intracellular and occurs specifically at the assembly stage [110]. It has been suggested that the timing of such initiation plays a role in maximising the efficiency with which budding occurs.

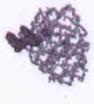


Figure 3.7: Electron micrograph image of a new HIV virion budding from a host cell (reproduced from: http://www.rhodes.edu/biology/glindquester/viruses/pagespass/hiv/hiv.html).

3.3 The Approach Towards HIV Treatment

An understanding of the life-cycle of HIV, elaborated on in §3.2, has opened up several avenues which have led to the design of anti-retroviral inhibitors (ARVs) of HIV. All of these classes of inhibitors seek to interfere with the life-cycle at crucial points. To this effect there are four classes of ARV, fusion inhibitors (FIs), nucleoside reverse transcriptase inhibitors (NRTIs), non-nucleoside reverse transcriptase inhibitors (NNRTIs) and protease inhibitors (PIs). Table 3.1 shows the ARVs approved by the US Food and Drug Administration (FDA)[3].

Nucleoside Reverse Transcriptase Inhibitors

NRTIs act as terminating nucleoside analogues that are inserted by RT when reverse transcribing viral RNA into DNA, preventing the completion of the viral DNA. Immature DNA termination results in the impeding of the HIV life-cycle prior to integration with the host genome. The first successful inhibitor developed against HIV was the NRTI, Zidovudine (3'-azido-3'-dexythymidine, AZT) [108].

Non-Nucleoside Reverse Transcriptase Inhibitors

NNRTIs work by binding directly to the RT itself and thus impeding its function. For example, Delavirdine, Efavirenz and Nevirapine bind to a region of RT distant from its catalytic site and so cause

[3] http://www.fda.gov

a conformational change in the active site that inhibits its activity [108].

Fusion Inhibitors

There is currently only one licensed fusion inhibitor, Enfurvitide (T20). In principle, fusion inhibitors work by interfering in the unique mechanism of viral entry into the host cell by association with the CD4 receptor and the CCR5 co-receptor. T20 works by immobilising gp41 in an intermediate structure incapable of triggering membrane fusion [108].

Protease Inhibitors

The protease forms a natural target for the inhibition of HIV due to the crucial role it plays in the maturation of the virus. Protease inhibitors (PIs) work by binding to the active site of the protease and thus blocking entry of the Gag and Gag-Pol polyprotein chains. We will discuss these in more detail in section §3.4.5.

The longstanding problem faced by most ARVs is the development of drug resistance by viral strains of HIV. The origin of drug resistance lies in the infidelity of the reverse transcription of RNA into DNA. Approximately one transcription error is made every 30000 base pairs, resulting in about 1 in 3 virion copies containing a mutation [111]. This, coupled to the large replication rate [112] of the virus ($\sim 10^8 - 10^9$ per day), leads to a vast rate of mutation in its genome. Most of these mutations will have insignificant or negative effects on the efficacy and function of the corresponding proteins but some will enhance the function of those proteins. The problem with treatment with ARVs is that although they

NRTI	NNRTI	PI	FI
Zidovudine (AZT)	Delavirdine	Saquinavir	Enfurvitide (T20)
Abacavir	Efavirenz	(Fos)Amprenavir	
Lamivudine	Nevirapine	Indinavir	
Stavudine		Lopinavir	
Didanosine		Nelfinavir	
Zalcitabine		Ritonavir	
Emtricitabine		Tipranavir	
Tenofovir		Atazanavir	
		Darunavir	

Table 3.1: Current FDA-approved anti-retroviral (ARV) inhibitors of HIV.

inhibit the proteins they are designed for, they place a selection pressure on the multiple strains of viral proteins that exist in the host. As a consequence, drug resistant mutant strains begin to proliferate and become dominant, leading to eventual failure of the original treatment. This has especially been the case with ARVs used singularly [108]. The development of drug resistance in the protease in particular is discussed in more detail in §3.4.6.

A milestone in the treatment of HIV was therefore the development of regimens involving multiple drug treatment. This has been characterised by the highly active anti-retroviral therapies (HAART) which use different combinations of multiple ARVs containing NRTIs, NNRTIs and PIs. HAART cocktails can often reduce viral counts to undetectable levels [108]. However the emerging problems with these cocktails is the strictness of the required regimens, the consequence of failure to adhere to these regimens being the rapid development of drug resistance and adverse side-effects [108]. The cocktails administered have also always got to keep evolving, to keep up with the continuous mutation of the virus. There is still therefore a need to develop optimal regimens as well as more specific and successful inhibitors of HIV.

3.4 The HIV Protease

3.4.1 Structure

HIV protease has been extensively studied, both experimentally and computationally, by a range of methods. At the time of writing, over 240 structures of the enzyme existed in the Protein Data Bank [11]. By far the great majority of these are crystallographic structures, although more recently a handful of NMR structures have been reported.

HIV protease is a homodimeric aspartyl protease [113], made from two identical domains each with a length of 99 amino acid residues. Figure 3.8 shows a schematic 'ribbon' diagram of the backbone structure of the protease. The enzyme has rotational C_2-symmetry about a central axis through the dimerisation interface.

The active site itself is formed by the dimer interface and is the region in which the catalytic process of the enzyme is carried out. The dimerisation interface is largely made up of a β-sheeted structure containing the first and last ten residues of each monomer and a central region from approximately residues 20 to 30 on each monomer, each of which contain a catalytically active aspartic acid (residue 25) forming the largely hydrophilic base of the active site. The top of the active site is enclosed by a pair of highly flexible hairpin β-sheets, again one from each monomer, known as the 'flaps' (residues 43-58), which, when in a closed conformation, complete the dimerisation interface at the flap tips (residues 48 to 52).

We will adopt the terminology of Perryman *et al.* (2004) in describing the structural components of

the protease after its resemblance to a bulldog's face [114]. The active site 'flaps' are connected to the 'ears', also known as the 'elbows' (residues 35-42), the 'cheek turn' or 'fulcrum' (residues 11-22), the 'cheek sheet' or 'cantilever' (residues 59-75), the eyes (residues 23-30) forming the base of the active site, the nose (residues 6-10) and the whiskers (residues 1-5) that form part of the dimerisation interface.

3.4.2 Function

As mentioned in §3.2.2, the protease, which unlike some other aspartic proteases, is only active in dimeric form, is responsible for proteolytic cleavage of the Gag and Gag-Pol polyprotein chains necessary for subsequent maturation of infectious virions. The protease recognises specific amino acid sequences and cleaves the peptide bond at 10 distinct cleavage sites along the Gag and Gag-Pol chains (see Figure 3.9). This is done by the catalytic aspartic acids (Asp) at residue positions 25 on each monomer, which sit at the base of the active site.

Gag and Gag-Pol are synthesised in a ratio of about 8:1 due to a -1 frameshift at the amino terminus of the p1 protein [115] preventing the termination of Gag and leading into Pol translation. Therefore, in addition to the structural proteins, MA, CA, p2, NC, p1 and p6 and the enzymatic proteins, PR, RT, RH and IN, a trans-frame region (TFR), consisting of overlapping amino acids in the Gag and Gag-Pol reading frames, is also generated. Due to several inconsistent naming schemes in the literature and the relevance to the following discussion, we provide a description of the TFR as adopted by Chatterjee *et al.* [116] that will be used henceforth. The TFR consists of a short trans-frame peptide (TFP), 8 amino acids long, as well as a variable longer region of between 48 - 60 amino acids, denoted as $p6_{Pol}$. Subsequently, neither the p1 nor the $p6_{Gag}$ proteins are generated from Gag-Pol translation and instead NC translation extends directly into TFP without subsequent cleavage, followed by $p6_{Pol}$, where the TFP-$p6_{Pol}$ site is subsequently cleaved [117]. The NC released from the Gag-Pol precursor is therefore actually NC-TFP and is 8 amino acids longer than NC released from Gag.

The full process of Gag and Gag-Pol cleavage by the protease is not fully understood. However, it is well known that alongside the protease only being active in the dimeric form, processing of both Gag and Gag-Pol precursors is achieved by the proteases encoded within the Gag-Pol chains. For such autoprocessing to begin, two Gag-Pol chains must therefore first partially dimerise, through which folding of an active dimeric protease can occur [118, 119]. Such folding has been shown to help initiate the autoprocessing of the Gag-Pol precursor, even though the free protease folds more completely than one embedded in a precursor [116].

The pseudo-folded and embedded protease can subsequently cleave either its own Gag-Pol chain or other Gag or Gag-Pol chains within the newly formed virion, although recent studies have shown that it first processes cleavage sites of its own Gag-Pol chain [119]. The order of Gag-Pol processing is also not fully understood. However, it has been established that the p2-NC site is the fastest cleaved, followed by the TFP-$p6_{Pol}$ site and that the N- terminus proline of the protease embedded in Gag-Pol is

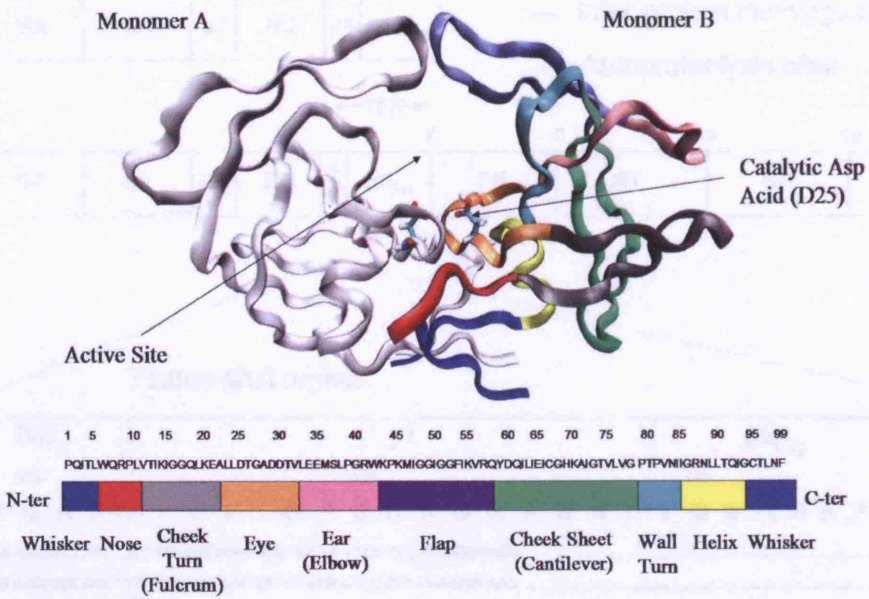


Figure 3.8: The three-dimensional structure of the amino acid backbone of HIV protease. Monomers A and B, each 99 amino acids long, dimerise to give a C_2 rotationally symmetric structure akin to a 'bulldog's face' [114]. The distinct regions of monomer B are colour coded. The active site is formed by the dimerisation process and enclosed by the 'flap' (43-58) of each monomer on top and the 'eye' (23-32) that contains the catalytic aspartic acid (D25) residue at its base. The N- and C- terminal 'whiskers' (1-5 and 95 -99) of each monomer interweave into a four stranded β-sheeted motif that forms a part of the dimerisation interface. The 'nose' (6-10) contains the residue R8 which covers the entrance to the front of the active site and leads into the 'cheek turn' or 'fulcrum' (11-22). The ears (33-42) are joined to the flaps and lead into the 'cheek sheet' or 'cantilever' (59-79). The 'wall turn' (79-85) forms part of the active site and leads into the α-helix structure (86-94) that supports the active site base. The clinical consensus wildtype sequence HXB2 is also shown.

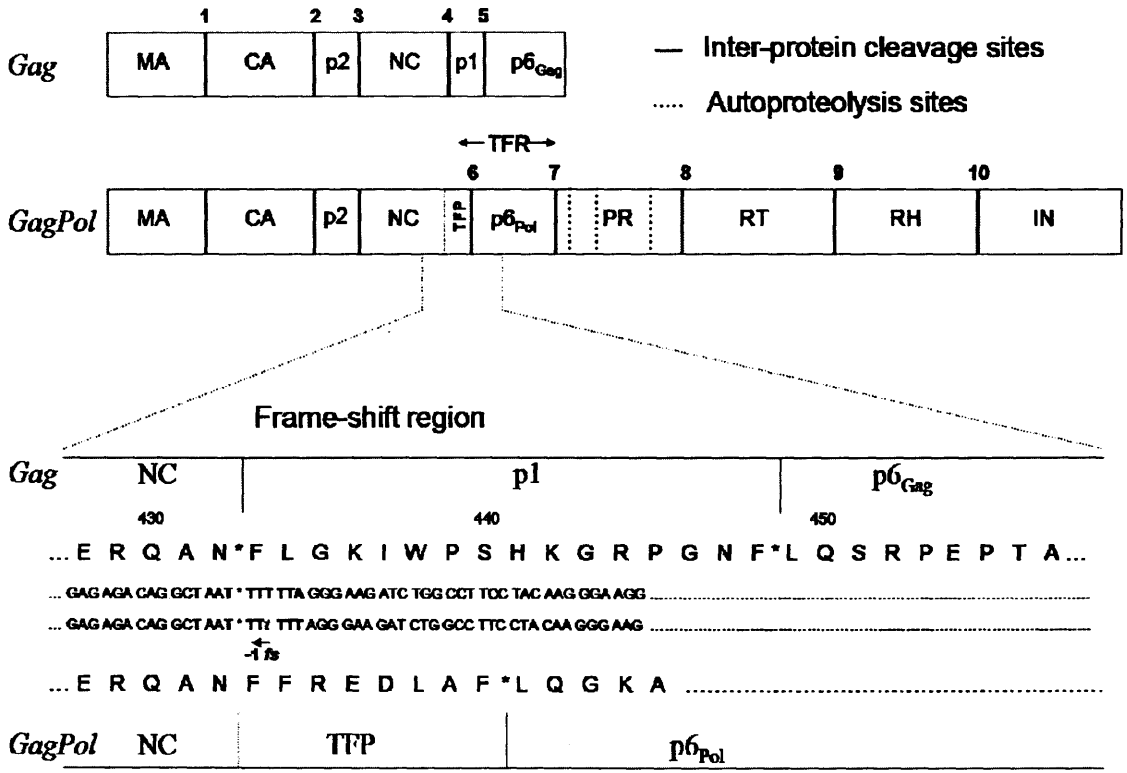


Figure 3.9: Constituents of the Gag and Gag-Pol precursors. Gag consists of the matrix (MA), capsid (CA), p2, nucleocapsid (NC), p1 and $p6_{Gag}$ proteins. A -1 frameshift (fs) whilst reading the RNA at F433 (TT*t*) prevents Gag termination and results in Gag-Pol with an overlapping section known as the trans-frame region (TFR). NC is therefore extended by a short 8 amino acid long trans-frame peptide (TFP), whilst $p6_{pol}$, protease (PR), reverse transcriptase (RT and RH) and integrase (IN) are translated too. There are therefore ten distinct inter protein cleavage sites (labelled 1-10) as well as three sites at which autoproteolysis occurs.

critical in directing the order of cleavage [119, 120].

Furthermore, such studies have demonstrated that the catalytic efficiency of Gag-Pol embedded protease is still significant, albeit not comparable to that of free protease. This supports previous studies that have also shown that the protease is catalytically active whilst being part of a $p6_{Pol}$-PR fusion protein [121]. Interestingly, proteases with the N-terminus free but with the C-terminus still bound to part of the precursor have been shown not to decrease catalytic efficiency by a significant amount [122].

By contrast, the order of cleavage of the Gag chain has been more thoroughly established and is as follows: p2/NC, MA/CA and p1/p6, NC/P1 and CA/p2 [123, 124]. From this it follows that the NC/p1 and CA/p2 are important as rate determining steps of Gag processing and subsequent viral maturation. Even though CA/p2 cleavage is slower *in vivo*, it stands out amongst the other sites in Gag, as it is the only one which is regulated by more than the interaction with the protease active site. *In vitro* studies, have shown that CA/p2 cleavage in the absence of the entire Gag precursor is comparable to MA/CA cleavage [125].

As well as the cleavage sites shown in Figure 3.9, the protease also degrades via auto-proteolysis at several points in its own sequence, typically Leu5-Trp6, Leu33-Glu34 and Leu63-Ile64 [126, 127]. Such inactivation is consistent with a model in which the dimeric protease is in equilibrium with disordered monomeric chains which, due to their unfolded nature, subsequently serve as cleavable substrates. Indeed, the fast rate of autoproteolysis has been a significant obstacle to the preservation of proteases prepared for subsequent study. Overcoming this obstacle has involved incorporating mutations at a combination of sites, specifically Q7K, L33I and L63I (see §3.4.6 for notation), which reduce degradation and are thus characteristic of many existing crystal structures.

Experimental studies usually involve peptide chains that match the sequence of the cleavage sites, often referred to as 'natural substrates'. Substrate specificity is important for the protease to distinguish the correct set of amino acid sequences to cleave and is mediated by hydrophobic and hydrophilic accommodation of side-chains of natural substrates into sub-pockets of the protease active site. Conventionally, the side chains of amino acid residues towards the N- terminal of the substrate peptide chain are designated as P1 - P4 outward of the cleaved peptide bond, whilst those towards the C- terminal are designated P1′ - P4′. These correspond to the S1 - S4 and S1′ - S4′ pockets in the protease active site respectively. Such nomenclature is also followed by the inhibitors of HIV protease, for which the P-subsites on the drug molecule are designed to target the corresponding S- pockets of the protease.

Table 3.2 shows the wildtype sequence of each of the natural substrates. Whilst no two substrates are identical and in fact contain significant variation, there are nonetheless some conserved properties that define the specificity of the cleavage process. The P1 and P1′ subsites are largely hydrophobic; P1 in particular is predominantly phenylalanine and is also the least varying of all cleavage sites. Furthermore, asparagine is found at position P2 in four cleavage sites, which is only slightly more varying than P1. It is however not surprising that there is a degree of variation in the peptide sequences in the context of

the above-mentioned difference in cleavage rates for different substrates.

It has conventionally been impossible to obtain crystal structures of HIV-1 protease bound to its natural substrates due to the susceptibility to cleavage of the peptide bond before crystallisation. Structural information with regard to substrate specificity has therefore been inferred from peptidomimetic inhibitors that are designed to imitate natural substrates, but crucially do not have a cleavable peptide bond. These are described in more detail in §3.4.5. Recently, however, a set of crystal structures were reported for several of the natural substrates bound to wildtype and mutant forms of the D25N inactive protease [128–131]. The conversion of the aspartic acid dyad to sterically similar asparagine allows for structural association between substrate and protease to be maintained, but importantly, prevents peptide cleavage, allowing for crystallisation of the complex.

3.4.3 Flexibility and Dynamics

Previous studies of the many crystal structures reported for both HIV protease and its complexes have revealed the characteristic flexibility of the flaps together with the stability of the catalytic region [132] (see Figure 3.10). In these studies 73 complexes of HIV-1 protease bound to various ligands were used to calculate the mean pairwise root mean squared deviation (RMSD) values of the protease backbone atoms (N, C_α, C). Figure 3.10 shows the same backbone RMSD but for a larger number of crystal structures (162), owing to the increased number of depositions in the Protein Data Bank and incorporating the backbone carbonyl oxygen atom (O) into the analysis. Alongside the flaps, the ears are also very flexible.

Substrates	**P5**	**P4**	**P3**	**P2**	**P1**	-	**P1′**	**P2′**	**P3′**	**P4′**	**P5′**
MA-CA	Val	Ser	Gln	Asn	Tyr	-	Pro	Ile	Val	Gln	Asn
CA-p2	Lys	Ala	Arg	Val	Leu	-	Ala	Glu	Ala	Met	Ser
p2-NC	Pro	Ala	Thr	Ile	Met	-	Met	Gln	Arg	Gly	Asn
NC-p1	Glu	Arg	Gln	Ala	Asn	-	Phe	Leu	Gly	Lys	Ile
p1-p6$_{Gag}$	Arg	Pro	Gly	Asn	Phe	-	Leu	Gln	Ser	Arg	Pro
TFP-p6$_{Pol}$	Glu	Asp	Leu	Ala	Phe	-	Leu	Gln	Gly	Lys	Ala
p6$_{Pol}$-PR	Val	Ser	Phe	Asn	Phe	-	Pro	Gln	Ile	Thr	Leu
PR-RT	Cys	Thr	Leu	Asn	Phe	-	Pro	Ile	Ser	Pro	Ile
RT-RH	Gly	Ala	Glu	Thr	Phe	-	Tyr	Val	Asp	Gly	Ala
RH-IN	Ile	Arg	Lys	Ile	Leu	-	Phe	Leu	Asp	Gly	Ile
AutoPR-a	Pro	Gln	Ile	Thr	Leu	-	Trp	Lys	Arg	Pro	Leu
AutoPR-b	Asp	Asp	Thr	Val	Leu	-	Glu	Glu	Met	Asn	Leu
AutoPR-c	Tyr	Asp	Gln	Ile	Leu	-	Ile	Glu	Ile	Cys	Gly

Table 3.2: Sequence specificity of protease substrate cleavage sites.

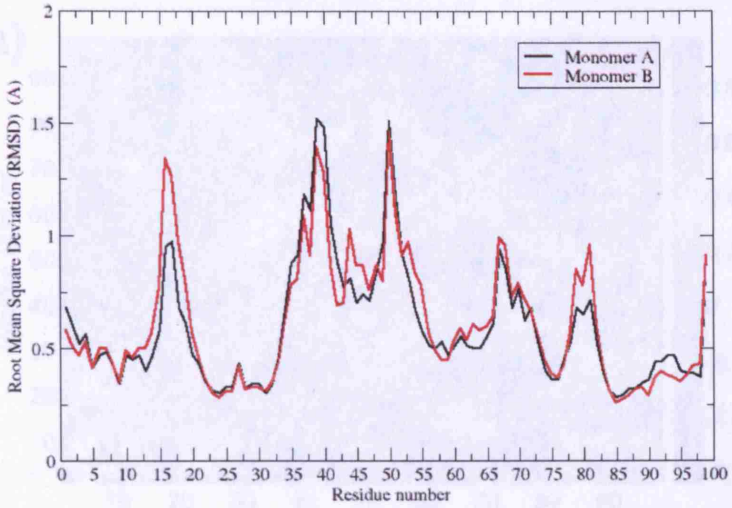


Figure 3.10: The mean pairwise RMSD from a set of 162 crystal structures across the 99 backbone residues of HIV-1 protease. Monomer A is shown in black whilst monomer B is in red. Whilst at the time of analysis there were over 230 structures in the PDB, not all of them are suitable for such an analysis due to the absence of resolved backbone atoms. The set chosen therefore consists of structures which have no such missing atoms.

There is a reduction in the flexibility in the flap region (43-48), likely due to the fact that most crystal structures have an inhibitor bound, for which underside flap motion is significantly decreased due to binding. The tips of the cheek turn, the cheek sheet and the wall turn display similar flexibility to the underside of the flaps, whilst the whiskers, the nose and expectedly the helix are inflexible (see Figure 3.8). It is an interesting characteristic that the most stable region, the eyes, containing the catalytic dyad, is not only the most inflexible region but is flanked sequentially by flexible regions. Furthermore, from a structural perspective, apart from the eyes, the rest of the active site region, namely the wall turn and the flaps are comparatively flexible. Such profiles of flexibility further provide insight into the mechanisms of substrate binding. The active site being the most stable must secure a peptide bond for cleavage, the flaps must be mobile to allow the substrate access and the wall turn must be optimised between sufficient flexibility to allow the conformational rearrangements of side-chains upon substrate binding whilst not being so flexible that such binding is itself hindered.

Alongside RMSD analysis, previous studies have also provided cross-correlation maps (CCMs) of the protease chains [132] from pairwise analysis of crystal structure data as well as molecular simulation [133]. CCMs provide insight into the conformational fluctuations of the protease and pairwise correlated motions of different parts of the protease. Interestingly however, whilst each monomer has been compared with itself, no data has been provided regarding the correlation of the monomers with each other. We therefore provide CCMs, constructed from the mean pairwise deviation of the backbone

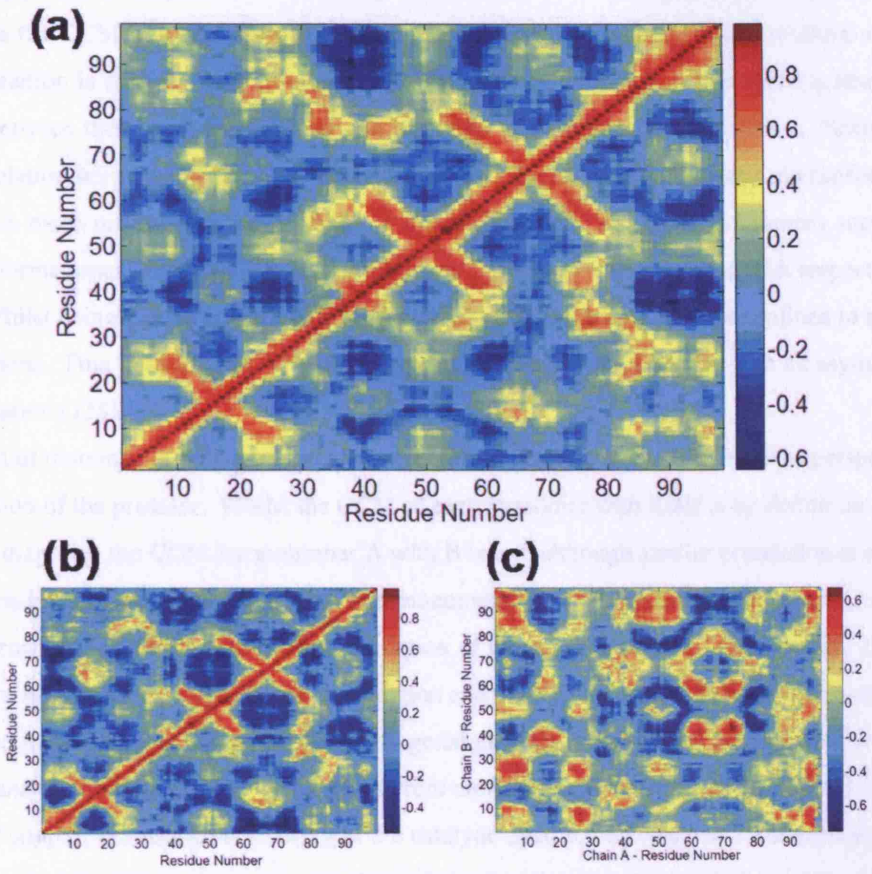


Figure 3.11: PDB-based cross correlation maps (CCMs) of HIV-1 protease. Positional correlations were calculated for the amino acid residue backbones from the pairwise RMSD of the 162 structures selected. (a) CCM of protease chain A with itself. (b) CCM of chain B with itself. (c) CCM of chain A with chain B.

atoms of each amino acid from the same set of 162 structures used in the RMSD analysis above, for each monomer with itself as well as for the correlation of monomer A with monomer B (see Figure 3.11). The specific details of calculating cross-correlations are provided in Chapter 2.

Our analysis of the correlation of each monomer with itself (Figure 3.11(a) and (b)) agrees well with that of previous studies [132]. In our analysis both correlations and anti-correlations for each monomer are presented in one half of the matrix, the other half being a mirror image in the diagonal. Apart from the expected diagonalised correlation, secondary structural features such as β-sheeted regions are represented by high correlation regions at normals to the diagonal, whereas the broadening of the diagonal in the 90s region represents the α-helix. Furthermore, there is strong correlation between the 'fulcrum' and the 'cantilever' as well as significant anti-correlation between the 'ear' and the C-terminal 'whiskers' and the ear with the helix region.

Interestingly, as our analysis does not average the effects of both monomers, we are able to observe differences in the CCM between monomer A and monomer B. For example, the C-terminal whisker-ear anti-correlation is slightly more pronounced in monomer A than in B, whilst there is strong anti-correlation between the N-terminal whisker and cantilever in monomer B and not in A. Furthermore, the anti-correlation between the flap and the cantilever and the correlation between the cantilever and the fulcrum is more pronounced in monomer B than in monomer A. These differences support the fact that conformational fluctuations in the protease are somewhat heterogeneous with respect to each monomer. Whilst being predominantly structurally symmetric, the protease is not confined to symmetrical fluctuations. This is supported by a recent crystal structure of the protease with an asymmetrical flap conformation [134].

The CCM of monomer A with monomer B (see Figure 3.11(c)) is interesting from the perspective of the dimerisation of the protease. Whilst the CCM of each monomer with itself is by definition a mirror image in the diagonal, the CCM for monomer A with B is not, although similar correlation is observed between corresponding amino acids in opposite monomers. The strong correlation in the corners of the CCM correspond to the C- and N-terminal region of the dimer interface. Additionally, along the diagonal there is strong correlation in the 'eye' region and the flap-tip region, both of which are part of the dimer interface. The correlations in the off-diagonal elements between the fulcrum and the N- and C-terminals and the fulcrum and the 'nose' of different monomers complete the description of the dimer interface and support the structural stability of the catalytic dyad region (residue 25 of each monomer) afforded through dimerisation. Interestingly, although the flap-tips are correlated, the greater part of the flaps themselves are anti-correlated and such opposing fluctuation supports the opening and closing role of the flaps. Other regions of significant correlation or anti-correlation include opposing C-terminal-flap regions, cantilever-ear regions, cantilever-cantilever regions and wall-flap regions.

The characteristic mobilities of the flaps and the catalytic site provide insight into the dynamical function of the protease and suggest several conformations that can exist in the flaps. In order for viral polyproteins to access the active site, it is intuitive to assume that the flaps open up to facilitate such a process and subsequently close around the ligand to secure binding. Indeed, whilst crystal structures of ligands bound to HIV-1 protease have shown the flaps in a 'closed' conformation in which the flap tips overlap (see Figure 3.12), in structures of the unliganded protease such as 1HHP, the conformation of the flaps is 'semi-open' and the flap tips no longer overlap, whilst maintaining a close proximity to each other. Although these definitions are somewhat arbitrary, previous studies have used the 1HHP structure as a definition of 'semi-open' [114]. There is currently only one crystal structure with the flaps in a completely 'open' conformation (pdb:1TW7) and this structure itself has been recently shown to be stabilised due to crystal packing effects [135].

Nonetheless, protease flexibility, stability and flap motion have also been extensively studied both experimentally and using computational techniques such as molecular dynamics [136–140]. These

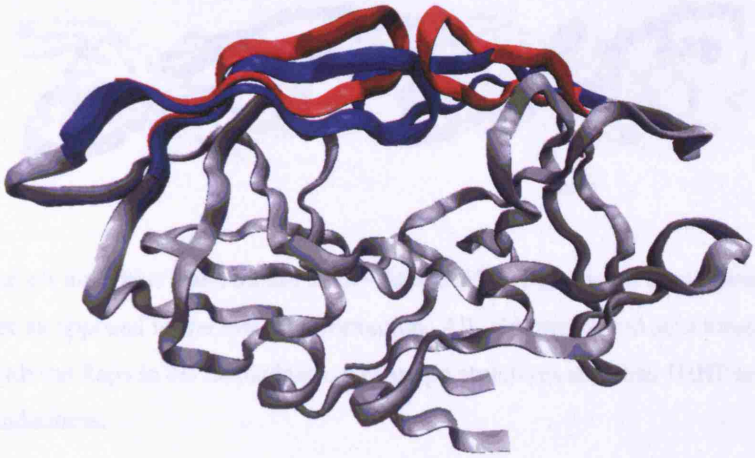


Figure 3.12: The 'semi-open' (red) and 'closed' (blue) conformations of the flaps of HIV protease, as represented in PDB structures 1HHP and 1HXB respectively. The inhibitor has been removed from the 1HXB structure for clarity. The flaps must open to facilitate access of viral polyproteins to the active site; crystal structures of the apo-protein, however, show a semi-open conformation, whilst ligand-bound structures are found in the closed conformation.

studies have shown that the flaps oscillate between various conformations, ranging from 'closed' to 'open' and that such motion modulates the access to the active site [141]. Recent studies on both the free and inhibitor-bound enzyme suggest the predominance of the semi-open flap conformation in the unliganded protease whilst showing that the flaps remain stably closed when an inhibitor is bound [142]. This is in good agreement with previous NMR studies [143, 144] that have suggested equilibria between open, semi-open and closed forms of the flaps.

The mechanisms which induce changes in flap conformation have also been studied. It has been reported by several authors that curling of the flap tips is related to subsequent changes in flap conformation from closed to open [114, 136, 145] and these observations are again consistent with NMR studies that show particularly significant motion of the flap tips (residues 49-53) [144].

Flap handedness is also an interesting feature of the protease. The orientation of the flaps in the semi-open conformation of 1HHP is opposite to that of the closed conformations in ligand bound structures (see Figure 3.13). We define the handedness observed in closed conformations as *cis*, as the 'flaps' do not cross over each other with respect to their pivot, and correspondingly the reverse handedness as *trans* in which such a cross-over is exhibited.

Recent studies have also shown that flap transition from closed through to semi-open also results in reversal of flap handedness prior to full flap opening [142] and it is interesting that such a transition results in the loss of hydrogen bonding between the flap tips. It has, however, not been established whether the reversal of handedness is required to reach a semi-open protease conformation or whether a semi-

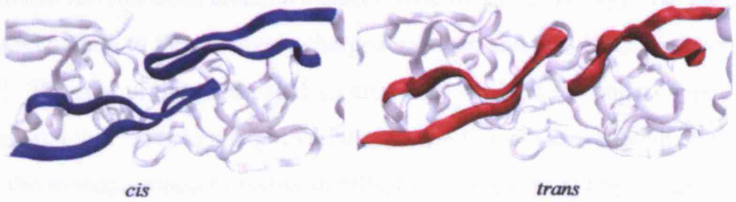


Figure 3.13: The *cis* and *trans* handedness of the flaps of HIV-1 protease. In *cis* handedness the flaps do not cross over as opposed to the *trans* conformation. All inhibitor-bound structures exist in a closed conformation with the flaps in *cis* handedness, whilst apo structures such has 1HHP are semi-open and exhibit *trans* handedness.

open conformation with handedness unchanged with respect to the closed conformation is possible. Interestingly, however, in the only crystal structure (1TW7) of the open conformation, the handedness of the flaps is again *cis* as compared to the *trans* conformation observed in the semi-open structure 1HHP. However, due to the above mentioned crystal packing effects associated with this structure, the handedness observed in 1TW7 in conjunction with its 'openness' may not be stable in solution as in molecular dynamics simulations it has been shown to revert to a semi-open conformation [135]. This, however, does not negate the viability of the 1TW7 structure as a possible open structure, as the open conformation observed by Hornak *et al.* [142] in MD simulations is also transient and itself reverts back to a semi-open conformation. Unfortunately, due to a combination of complexity in crystallographically obtaining an open structure as well as the computational demand in accessing timescales that lead to flap opening, the exact form of the flaps in an open conformation is not yet fully understood.

In Chapter 5, we report a study into the differential dynamics between wildtype and mutant forms of the HIV-1 protease when bound to an inhibitor and comment on the observed transition of the flaps in relation to the above discussion.

3.4.4 Enzymatic Mechanism

The enzymatic mechanism of peptide bond cleavage has been extensively studied over the past two decades, although it is still not fully understood [146]. One of the currently favoured mechanisms is the general acid/general base (GA/GB) mechanism which is supported by many studies [133, 147–153]. The GA/GB mechanism is shown in Figure 3.14 and involves a lytic water molecule, one involved in the cleavage process, as well as the requirement that one of the aspartic acid residues in the catalytic dyad be protonated. The water molecule, which is polarised by the Asp dyad, attacks the substrate carbonyl carbon atom forming a transition state in which the dissociated hydroxide group of the water binds to the peptidic carbonyl carbon. This state progresses to an intermediate state with the Asp dyad donating a proton to the peptidic nitrogen atom. The resulting transitional zwitterionic state of the

peptide group facilitates simultaneous electron transfer from the carbonyl oxygen to the carbonyl carbon and from the carbonyl carbon to the positively charged nitrogen and thus leads to decomposition of the peptide bond [147]. The role of the protonated aspartic acid varies according to where it is in relation to the cleaved group and this leads to a slight variation in the GA/GB mechanism. For example, in the above mechanism, the monoprotonation assists stabilisation of the aspartyl dyad through formation of a strong inter aspartyl hydrogen bond. In alternative formulations [133], monoprotonation of the aspartyl dyad actually facilitates destabilisation of the peptide bond through hydrogen bond formation with the carbonyl carbon.

An alternative mechanism that has been suggested is that of concerted nucleophilic attack on the carbonyl bond of the substrate cleavage site by the oxygen atoms of one of the aspartic acid residues [147, 154]. Such a mechanism (also shown in Figure 3.14) does not require a lytic water molecule to be present, nor does it place stringent requirements on the protonation state of the dyad. Instead, direct nucleophilic attack on the carbonyl carbon occurs via proton exchange from an aspartic acid to the peptide nitrogen. The other negatively charged aspartic acid then directly attacks the carbonyl carbon. In the example shown in Figure 3.14, the diprotonated aspartyl dyad loses a proton through the first stages of the reaction, being reduced to a monoprotonated state. Again, the second protonation stabilises the aspartyl pair. In the slightly varying mechanism proposed by Park *et al.* [154], the initial state is monoprotonated and the resulting cleavage process results in formation of a dianionic dyad. However, even though computational studies have shown that the nucleophilic mechanism is feasible [147, 154], there is currently no direct experimental evidence for it.

As the enzymatic mechanism is related to the protonation of the aspartyl dyad, it is not surprising that several studies have been conducted in this regard. Experimental studies have shown that the protease has a characteristic bell-shaped activity profile in the pH range 3.5 - 6.5 and is most effective at a pH ~5-6 [155]. For optimum activity therefore, the favoured state should be monoprotonated. However, there is much uncertainty and debate as to the actual protonation of the protease and a range of theoretical and experimental studies have shown the feasibility of the mono-, di- and unprotonated dyad states [138, 150, 151, 155–160]. Whilst at a physiological pH of 7.4 the catalytic dyad should be unprotonated, some studies also support a dianionic state in the apo-protease at the weakly acidic pH 6 [138, 160]. Furthermore, the influence of the protonation state upon ligand binding has also been extensively studied and several ligands have been shown to bind to varying protonation states [161–165]. Recent studies have even shown that ions can be a feasible substitute for protonation and stabilise the protease in the absence of ligands [156]. These studies suggest that subsequent substitution of a catalytic aspartyl-bound positive ion such as Na^+ for a proton can occur upon or after ligand binding.

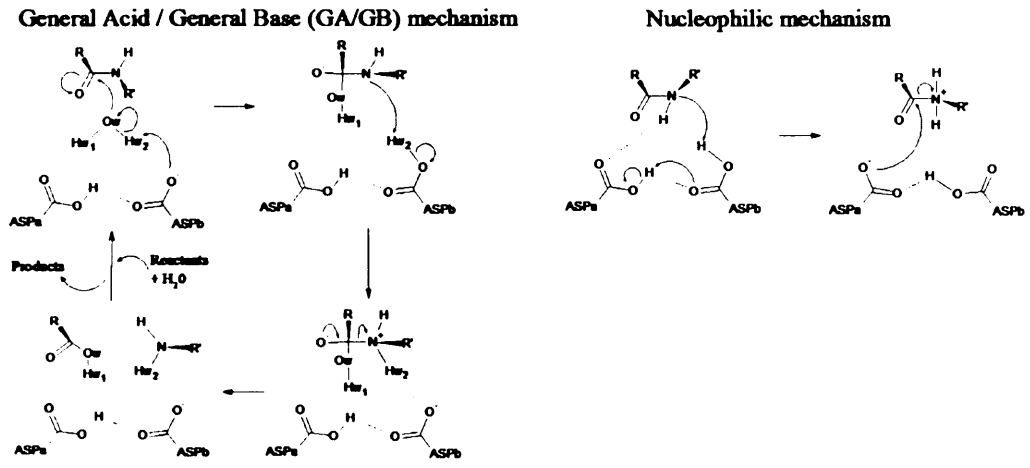


Figure 3.14: The two proposed mechanisms of enzymatic cleavage of the amino acid peptide bond by HIV-1 protease. The most favoured is the general acid/general base (GA/GB) type mechanism which involves a lytic water molecule and a monoprotonated aspartyl dyad. Hydrolysis of the peptide bond uses up the water molecule and so a new one as well as a new substrate are required to repeat the cycle. The alternative mechanism is a direct nucleophilic attack on the carbonyl carbon of the peptide bond (only the cleavage part of the process is shown).

3.4.5 Inhibitors of HIV-1 Protease

The large number of resolved crystallographic structures of HIV protease has led to the protease being a key example of rational (structure-assisted) drug design [166]. There are currently nine FDA approved inhibitors of HIV-1 protease (shown in Figure 3.15).

The design of HIV protease inhibitors has followed peptidomimetic principles, the first effective inhibitors being variants of peptide chains with particular side chain components that fitted in the binding sub-pockets (denoted S1 - S4 and S1′ - S4′) present in the active site. Like the natural substrates that bind to the protease, it is conventional to denote the side-chains of such inhibitors with the same nomenclature (P1 - P4 and P1′ - P4′) based on which active site sub-pocket they have been targeted for.

The first inhibitor of HIV-1 protease, saquinavir relied on the basic design criterion that the viral protease, unlike other proteases, is able to cut Tyr-Pro and Phe-Pro sequences in the viral polyprotein [166]. Furthermore, a crucial difference to a conventional peptide was the replacement of the cleavable peptide bond by an uncleavable hydroxyethylene moiety that binds to the dyad. Unlike saquinavir, which utilised principles of asymmetry in its design, drugs developed later, exploited the idea of using a largely symmetrical molecule within the protease. This led to the design of ritonavir [166].

The crystal structures of HIV protease/inhibitor complexes have revealed a conserved water molecule that is tetrahedrally coordinated between the flaps and the inhibitor. This water molecule is hydrogen-

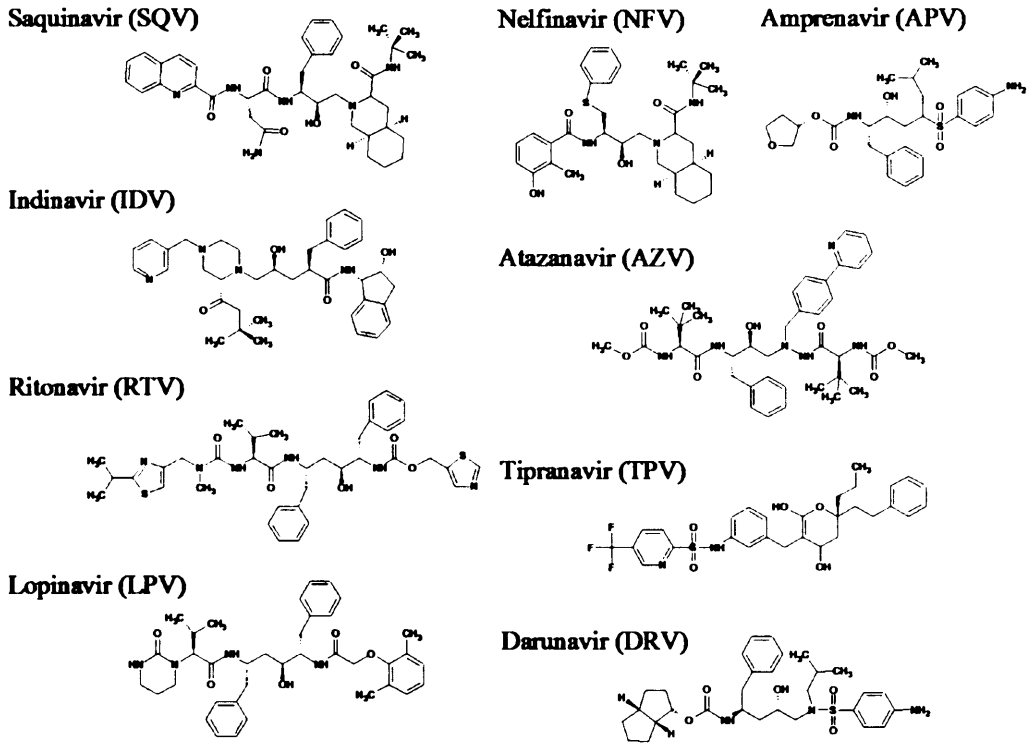


Figure 3.15: The nine FDA approved inhibitors of HIV-1 protease.

bonded, accepting two hydrogens from the isoleucine 50 residues of both monomers and donating two hydrogens to the carbonyl oxygen atoms of the inhibitor [22]. This has led to the suggestion of designing inhibitors that replace the water molecule, providing added specificity to binding as well as being more favourable due to entropic increase through the displacement of the original water molecule. An implementation of this suggestion was carried out by Lam *et al.* (1994) at DuPont Merck [167]. They designed an inhibitor that used a central cyclic urea with an oxygen atom that replaced the function of the original water molecule. Unfortunately, as of yet no cyclic urea based designs have been approved by the FDA, but several have been under clinical trials.

Over the last two decades, extensive experimental work has been conducted to determine and enhance the potency of a large array of inhibitors through the evaluation of the strength with which inhibitors bind to the protease. Enzyme Inhibition Assays (EIA) and Isothermal Titration Calorimetry (ITC), discussed in Chapter 1, have been the main techniques employed for the determination of inhibitor binding affinities, which are provided in terms of inhibition constants (K_i) and/or directly in terms of the free energy differences of ligand binding (ΔG_b). Also common are assessments of IC_{50}, the concentration of inhibitor required to halve the enzymatic activity. Much of this accumulated wealth of data has been deposited in a web-based database known as 'BindingDb' [16, 17]. Table 3.3 shows the potency of the current FDA-approved inhibitors with the wildtype HIV-1 protease, extracted from

BindingDb.

Inhibitor	K_i (nM)	ΔG (kcal/mol)	pH	Temp (°C)
Saquinavir	0.04	-14.19	6.4	25
Indinavir	0.07	-13.86	6.4	25
Ritonavir	0.02	-14.60	6.4	25
Nelfinavir	0.01	-15.02	6.4	25
Amprenavir	0.04	-14.19	6.4	25
Lopinavir	0.005	-15.69	4.7	30
Atazanavir	0.48	-13.23	4.7	37
Tipranavir	0.008	-15.00	5	22
Darunavir	0.014	-14.82	6.4	25

Table 3.3: Binding affinities of HIV-1 protease inhibitors with wildtype protease. Values extracted from the BindingDb database (see http://www.bindingdb.org) [17].

Whilst early inhibitors, such as saquinavir, with potency on the sub-nM level were initially successful, the emergence of drug resistant mutations in the viral protease (discussed in §3.4.6) has led to the evolution of strategies in drug design and the necessity to experimentally design ever more potent inhibitors, such as lopinavir, that approach the sub-pM level. Enhancement of inhibitor potency has also involved optimisation of the binding characteristics of the side-chain subsites of inhibitors [168]. However, whilst designing inhibitors with increasing potency is desirable, it is by no means the sole metric of an inhibitor's effectiveness *in vivo*. Alongside potency, it is important that inhibitor selectivity is maintained so that adverse side effects resulting from an inhibitor binding to non-desired human protein targets are prevented. Such toxic effects are usually tested experimentally in clinical trials but it is not uncommon for FDA approved inhibitors of the protease to have toxic side effects. Furthermore, the efficacy of a drug *in vivo* is governed by a number of additional factors such as its solubility as well as the ease by which it can reach its target. Therefore, inhibitors of HIV protease that have successfully emerged from clinical trials have had to optimise a range of pharmaceutical conditions.

3.4.6 Development of Drug Resistance in HIV-1 Protease

As discussed briefly in §3.3, treatment with anti-retroviral inhibitors has led to the emergence of drug resistant mutant strains that eventually render treatment ineffective. As this thesis investigates the effects of drug resistant mutations of the HIV-1 protease, here we provide an overview of some of the mechanisms through which mutations related to the HIV-1 protease can confer drug resistance as well as looking at some of the mutational patterns that are emerging from the treatment of specific protease inhibitors. The subject matter of later chapters, specifically Chapters 4-7, is concerned with the investi-

gation of some of these mechanisms from the perspective of molecular dynamics simulations and will be elaborated upon therein.

Mutations in the amino acid sequence of the protease are inevitable due to the infidelity of the reverse transcription process. Therefore, in any infected individual there is a 'swarm' of varying mutational strains and natural selection ensures that only strains that are 'fit' enough to process polyprotein precursors sufficiently well survive within the population. This sets immediate constraints on the genetic variability of the protease and it is not surprising therefore that several positions along the amino acid chain are conserved. For example, the active site triad (Asp25-Thr26-Gly27), which is crucial for the catalytic mechanism, is always conserved; high conservation is also observed at the N- and C- termini, the dimer interface and the substrate binding region [169].

However, notwithstanding such constraints, clinical studies have shown the protease to have a significant number of mutational strains even in untreated individuals [170]. This is consistent with studies on the three dimensional structure of the protease which exhibits remarkable tolerance to mutations [132]. In fact, up to half of the positions in the 99 amino acid chain have been shown to tolerate mutation, whilst maintaining a functional three dimensional structure. This is consistent with the genetic difference between HIV-1 and HIV-2 which is approximately 45 - 50 mutations in the protease. This serves as an example of the diversity of sequence-structure relationships in proteins. Many other proteins, in contrast, such as p53 (the protein involved in tumour suppression), show little or no tolerance to mutation and single point mutations can cause large structural and functional changes which render the protein inactive [171].

Such sustainable mutability in HIV protease therefore implies that many mutations have a negligible or small effect on the catalytic efficiency of the protease, whilst others severely diminish protease function. Within such a context, drug-associated mutations observed in clinical studies are reported on the basis of statistical significance with respect to treatment-naive patients. There have been many such clinical studies on a whole range of protease inhibitors [170, 172–178], from which insights into the mutational pattern of drug resistance can be gleaned. For example, Wu *et al.* [170] have suggested 17 non-treatment related polymorphisms as well as 37 residues that rarely if ever mutate, whilst Schinazi *et al.* [172] have shown there to be exactly 45 treatment-associated mutational positions.

Unlike reverse transcriptase, where single drug-associated mutations have been shown to cause significant resistance [173], the case for the protease is more complex. Several studies have been conducted showing how the protease accumulates mutations in an ordered and varied fashion, in response to inhibitor treatment [179–182]. Alongside key primary mutations that correlate strongly with treatment failure, there are patterns corresponding to the build up of several accessory mutations (sometimes up to 7 or 8), for which significant resistance may only be exhibited once several of these mutations are present in the individual.

Table 3.4 shows the characteristic mutations that have developed in response to treatment with key

protease inhibitors. Due to the large rate of viral mutation, it is not surprising that mutations have arisen in response to many of the early inhibitors of the protease [183]. Resistance mutations have even emerged for recently developed inhibitors such as tipranavir [184, 185] and darunavir (TMC-114), which studies have shown to be highly active against former drug resistant strains [186, 187]. Whilst it has been conventional to assume a deterministic model for the evolution of HIV-1 due to a large viral population size, stochastic models that allow chance to influence the direction of mutations in response to sub-optimal treatment, based on a smaller population size, have also been considered [188]. The effects of switching inhibitors after a mutational pattern has evolved, in order to determine the efficacy of salvage therapy, have also been studied, again showing the extensive malleability of the protease that permits it to adapt to whatever chemotherapeutic environment it is placed in [189]. Even more problematic however, is the emergence of strains that are multi-drug resistant (MDR), for which the efficacy of salvage therapy through the alteration of protease inhibitors is severely reduced [190].

As the protease is a homodimer, any single point mutation in the gene of the viral protease results in expression on both of its monomers. It is therefore convenient to adopt a notation that uniquely defines mutations and which is used extensively throughout this thesis. It is conventional to adopt a 'From residue type-Residue position number-To residue type' system when labelling mutations. For example the G48V mutation refers to the mutation of glycine to valine on the 48^{th} residue of each monomer. The residue position number starts from the N- terminus and finishes at the C- terminus. The residue type is expressed by a single conventional letter. Similarly, when describing the residues of the wildtype protease, it is conventional just to use the amino acid letter followed by the residue position number (e.g. I50 refers to isoleucine at position 50 of each monomer).

Resistance Mutations that Affect Drug-Protease Binding

Due to the need to overcome the limiting problem induced by drug resistance, the effect of mutations on the structural, dynamical and binding properties of HIV-1 protease with both inhibitors and natural substrates have been extensively studied. We will focus in more detail on some of these strategies in subsequent chapters. Here, it is fitting to outline in brief the general mechanisms through which mutations can confer chemotherapeutic resistance. The principal mechanism of resistance is for a mutation to reduce the binding affinity of a particular drug with the protease. This is consistent with the fact that several mutations such as V82A and I84V, occur in the active site (which normally, in untreated patients, is not susceptible to significant mutation) and thus exhibit a direct influence on inhibitor binding. More subtle mutations also arise away from the active site yet result in resistance to several inhibitors (L90M is an example of this). For such mutations, it is more difficult to explain the cause of the observed resistance.

Extensive experimental studies to determine the reduction in binding affinity for mutant proteases compared to the wildtype have been conducted for a significant array of mutants across a range of

Inhibitor	Associated Mutations
Saquinavir	L10I G**48**V I54V,L A71V,T G73S V77I V82A I84V L**90**M
Indinavir	L10I,R,V K20M,R L24I V32I M36I M46I,L I54V A71V,T G73S,A V77I V**82**A,F,T I**84**V L90M
Ritonavir	L10F,I,R,V K20M,R V32I L33F M36I M46I,L I50,V I54V,L A71V,T V77I V**82**A,F,T,S I**84**V L90M
Lopinavir	L10F,I,R,V K20M,R L24I V32I L33F M46I,L I**47**V,A I50V F53L I54V,L,A,M,T,S L63P A71V,T G73S V82A,F,T,S I84V L90M
Nelfinavir	L10F,I D**30**N M36I M46I,L A71V,T V77I V82A,F,T,S I84V N88D,S L**90**M
(Fos)Amprenavir	L10F,I,R,V V32I M46I,L I47V I**50**V I54L,V,M G73S V82A,F,S,T I**84**V L90M
Atazanavir	L10I,F,V K20R,M,I L24I V32I L33I,F M36I,L,V M46I G48V I**50**L I54L A71V G73C,S,T,A V82A I**84**V N88S L90M
Tipranavir	L10V I13V K20M,LT L**33**I,F E35G M36I K43T M46L I47V I54A,M,V Q58E H69K, T74P, V82L,T N83D I84V L90M
Darunavir	V11I V32I L33F I47V I50V L54L,V,M G73S L76V I84V L89V

Numbers in bold denote positions on each protease monomer unit strongly prone to mutation in response to the given inhibitor.

Table 3.4: Characteristic drug-associated mutations of HIV-1 protease [173, 191].

inhibitors using both enzyme inhibition assaying techniques [15, 192–196] and isothermal titration calorimetry (ITC) [197–202] (see Chapter 1). These reveal significant diversity in the way particular mutants confer resistance. Whilst some single or double mutations, such as G48V/L90M in response to saquinavir [15] or V82F/I84V in response to ritonavir [202], can cause more than 1000-fold drops in binding, studies have also shown severe drops in affinity from the accumulation of several accessory mutations, many of which are not in the active site [203] and which do not singularly cause significant resistance. Furthermore, studies on multi-drug resistant (MDR) mutants show that the acquisition of particular mutations within and away from the active site, can themselves cause over 1000-fold drops in affinity for several inhibitors [204]. These studies have also shown that significant cooperative interactions occur between particular sets of mutations, such that the presence of all mutations in a set confers more resistance than the sum of the contributions from the presence of individual mutations. Therefore, whilst primary mutations confer significant reduction in binding affinity, the accumulation of accessory mutants can further enhance the resistance profile [205, 206]. Interestingly, whilst mutations evolve to confer resistance to specific inhibitors, recent studies have shown that they can then become hypersusceptible to other inhibitors [207]. An example is the hypersusceptibility of I47A to saquinavir, which evolves in response to lopinavir treatment [208]. Studies have also shown that the observed changes in binding affinity between inhibitor and protease may, for some mutations, be due to an increase in dimer dissociation [209].

Studies on the decomposed binding energetics of inhibitors using ITC show that mutations can differentially alter the enthalpic and entropic contributions to binding [197, 198, 201] and that subsequent inhibitor design should factor in such capabilities. For example, the inhibitor KNI-764 exhibits greater binding affinity for the V82F/I84V mutant than first-generation inhibitors such as saquinavir, indinavir and nelfinavir show for the wildtype. Decomposition of the binding into enthalpic and entropic components have shown that, whilst there is an unfavourable change in enthalpy for the first-generation inhibitors binding to the mutant, the change in entropy for KNI-764 is favourable and thus opposed to the first-generation inhibitors which all show unfavourable entropy changes. Interestingly, the design of the recently developed inhibitor darunavir was motivated by the strong enthalpically driven binding observed for it [187]. Thus an improved understanding of the different components of binding should enhance the design of optimal inhibitors. Such inhibitors may not necessarily be those with the highest binding affinity, but instead those that optimise selectivity of the target with adaptability to evolving mutations of the protease [210].

Furthermore, other studies have shown that an inhibitor's torsional flexibility can be exploited to achieve binding to different mutational strains [211]. In Chapter 4 we discuss such torsional flexibility in the context of the multiple conformations observed in the inhibitor saquinavir with a range of mutants. Interestingly, recent studies showed the importance of considering an envelope constructed from the van der Waals surface of several superimposed substrates [212–215]. These studies showed significant

correlations between the sites of drug resistant mutations and the subsites of various inhibitors whose van der Waals surfaces extend beyond that of the natural substrates. Such a substrate 'envelope' thus provides a template for optimised drug design, an effective inhibitor being within the bounds of the envelope.

Whilst, experimental studies have provided a wealth of data on the effect of mutations on the drop in binding affinity and have even provided insight into enthalpic and entropic components of binding (see Chapter 1), investigating the kinetic mechanisms of drug resistance remain extremely difficult using current experimental techniques.

One strategy has therefore been to investigate the role of mutations in altering the dynamics of the flaps that modulate access to the active site as well as the conformational flexibility of the enzyme using molecular dynamics techniques [114, 139, 216]. The effect of mutations in altering the equilibrium between open and closed conformations of the flaps has been investigated. Early molecular dynamics simulations showed that the M46I mutation can cause the flaps to become stabilised in the closed conformation [217]. More recent studies have shown that the V82F/I84V mutational pair cause the flaps to sample more semi-open conformations than the wildtype [114]. Molecular dynamics techniques have also been used to investigate the differences in binding between mutant and wildtype proteases with several inhibitors [218] as well as in studies of dimer flexibility and stability [138, 219, 220]. However, as the subject matter of later chapters in this thesis is concerned with the application of molecular dynamics in understanding the effects of drug resistant mutations, we reserve further discussion of the topic until then (see Chapter 4-7).

Compensatory Mutations and the Alteration of Enzymatic and Viral Fitness

It is believed that the continual and rapid replication of HIV has allowed it to develop an optimal fitness *in vivo*, such that wildtype strains replicate faster than any other strains that have evolved. Several studies have shown that suboptimal antiviral therapy results in the selection of drug resistant mutants that are not as fit as the wildtype in the absence of chemotherapeutic pressure [221–223]. However, the selection of such mutants under antiviral therapy indicates that their fitness is then greater than that of the wildtype. This fitness inversion has its roots in the interplay between the effects of selected mutations on both the alteration of drug binding and the alteration of catalytic efficiency of the enzyme.

Several studies have shown that drug-associated mutations can cause deleterious effects on the catalytic efficiency of the protease through a reduction in the catalytic specificity rate k_{cat}/K_m [204, 223, 224] (see Chapter 1). Therefore, in the presence of an inhibitor, mutations that evolve need to optimise between the effects of reduced catalytic efficiency and reduced inhibitor binding. For such a scenario, only mutations whose inhibition reduction outweighs catalytic reduction will evolve in response to inhibitor treatment. In this context, studies have shown that compensatory mutations can arise, whose effect is to partially restore catalytic efficiency [225–227], and which serve as a secondary mechanism

by which mutations can confer resistance. For example, the mutations V82F and I84V show reductions in catalytic rate compared to the wildtype, whilst M46I in conjunction with V82F has a greater catalytic rate than V82F alone [224].

A combined description of the alterations in inhibitor binding and catalytic efficiency has resulted in the construction of a *vitality* metric (**V**) devised by Gulnik *et al.* [224] and used in similar forms in several studies since [203, 204, 206]. This is given as:

$$\mathbf{V} = \frac{(K_i.k_{cat}/K_m)_{mut}}{(K_i.k_{cat}/K_m)_{wt}} \tag{3.1}$$

where both the inhibition constant (K_i) and the catalytic specificity constant (k_{cat}/K_m) are combined to provide the vitality of a particular mutant (*mut*) with reference to the wildtype (*wt*). In this way, even though a mutation may decrease catalytic efficiency, provided it causes a large enough increase in drug dissociation its *vitality* will always be greater than that of the wildtype. By the same token, any subsequent compensatory mutations that cause increased catalytic efficiency but no significant increase in drug dissociation will also have a higher vitality. The metric is therefore an excellent way of ranking the enzymatic fitness of a particular mutant strain, whether the mutation induces a primary drug resistant effect or a compensatory effect through restoration of catalytic efficiency. Even mutations that induced greater drug binding, but whose catalytic efficiency increased sufficiently might be advantageous.

However, even though the vitality metric provides a good description of enzymatic fitness under drug pressure, it cannot be correlated directly to the overall fitness of the virus. The viral fitness is a measure of the capacity for viral replication both in the presence and absence of an inhibitor. This is governed by several parameters, including inhibitor concentration as well as the impact of mutations on the enzymatic performance of the protease [228]. Furthermore, an optimal viral fitness as observed for the wildtype does not necessarily correlate with optimal enzymatic performance. Studies have shown that mutations such as L90M can enhance the catalytic rate of the protease beyond the wildtype [15, 229, 230], yet their absence in the wildtype is indicative of reduced overall viral fitness when not under chemotherapeutic pressure. Attempts to describe viral fitness as a function of the enzymatic reaction rate have been made [231]. Unfortunately, these have been unable to describe the changes in fitness that occur when going from a drug-free to a drug-applied regime and as of yet no such theoretical framework exists. The additional complexity in developing such a framework is the fact that HIV exists not in singular strains, but as a quasi-species of many simultaneously existing strains of varying proportions [232]. A general framework would have to additionally consider how the fitness of a single strain relates to the fitness of the overall quasi-species in any infected individual. In Chapter 7 we investigate the effect of mutations on enzymatic fitness in more detail.

Mutations in the Gag Polyprotein

A third mechanism of resistance involves mutations not on the protease itself, but on the polyprotein chains that it processes. A comparatively recent set of mutations in some of the Gag precursor cleavage sites have been reported in association with protease inhibitor treatment [123, 233–236]. It is widely considered that such mutations are another form of compensatory mutation that evolve to enhance catalytic efficiency in response to primary drug resistant mutations on the protease. Unlike mutations on the protease, a significant way Gag mutations can confer resistance is through greater catalytic specificity rates (k_{cat}/K_m) than the wildtype. This is confirmed by several studies on cleavage-site mutations, both in the slowly cleaved NC-p1 site and the p1-p6 site which show enhanced catalytic specificity rates over the wildtype [229, 230, 237]. The absence of such mutations in untreated individuals who predominantly carry the wildtype again highlights the fact that increased catalytic rate does not necessarily correspond to increased overall viral fitness; only when the viral fitness of the wildtype is reduced under chemotherapeutic pressure can mutations that enhance catalytic efficiency increase the overall fitness profile.

One such example of a mutational pair involving cleavage site mutations is the V82A/A431V pair which has been observed in response to ritonavir. The A431V mutation corresponds to the P2 subsite of the NC-p1 cleavage site (see Table 3.2). Such studies are consistent with the observation that V82A itself causes resistance to ritonavir and recent structural studies have shown that V82A leads to a reduced hydrophobic interaction with the NC-p1 substrate subsequently restored by A431V [131]. Additional Gag mutations such as L449F and P453L, which occur in the p1-p6 cleavage site at subsites P1′ and P5′, have also been reported [238], as well as some non-cleavage site mutations [239, 240].

Finally, whilst not yet reported, it is plausible that Gag mutations may also arise as primary mutations in response to inhibitor treatment, solely in order to increase the catalytic efficiency of poly-protein processing and thus the subsequent viral fitness under chemotherapeutic pressure.

Aims and Objectives of Research

In the first three chapters of this thesis we have provided a review of some of the experimental and computational methods used to probe protein structure and dynamics, focussing in detail on the application of molecular simulation methodologies. Furthermore, we have provided an overview of HIV, concentrating on the structure, dynamics and function of its protease as well as the emergent phenomenon of drug resistance in response to anti-retroviral inhibitor therapy.

In the following four chapters we build upon the extensive body of research accummulated towards understanding the dynamics of HIV-1 protease. The central aim of our research is to further elucidate the phenomenon of drug resistance at the molecular level through the utilisation of molecular dynamics (MD) simulation methods. The research focusses on the molecular interaction of the wildtype HIV-1 protease and a set of three characteristic drug-resistant mutants, namely the G48V and L90M single mutants and the G48V/L90M double mutant with the inhibitor saquinavir.

The study reported in Chapter 4 aims to investigate the conformational flexibility of the inhibitor in the active site of the protease. Our objective is to determine whether a multitude of stable inhibitor conformations can exist and, if so, to describe the differential interactions between the inhibitor and the protease induced by such conformations. Furthermore, we investigate whether it is possible to distinguish mutation-dependent preferred conformational states.

The study presented in Chapter 5 builds directly upon that reported in Chapter 4. We aim to determine the longer timescale effects of differential drug binding in the active site, precipitated by the existence of the above-mentioned mutations.

In Chapter 6, we quantitatively investigate the thermodynamic basis of drug resistance. Our aim is to accurately calculate absolute binding free energies for the above protease variants with saquinavir using a combination of the MMPBSA and configurational entropy determination methods. Additionally, we aim to determine to what extent and accuracy the drug resistance conferred by the mutations with respect to the wildtype can be correctly ranked using our protocol. As the construction and execution of MD simulations is an involved and laborious task, a further objective has been to develop a tool for the automation of simulations as well as binding free energy calculations of HIV-1 protease-ligand variants. To this end, we report the development of the Binding Affinity Calculator (BAC), described more fully in Appendix A.

In Chapter 7, we focus our attention not on an inhibitor of the protease but on one of its naturally processed substrates. We aim to calculate the binding free energies of the above-mentioned protease variants with this natural substrate and thus provide insight into the effect of drug resistant mutations on the catalytic efficiency of the protease. A final objective is to determine how a suitable combination of the changes in both drug resistance and catalytic efficiency, both computable through molecular simulation, can provide a better description of the overall enzymatic fitness of a particular mutant strain of the protease in the presence of an inhibitor.

CHAPTER 4

Mutation-Altered Distribution of Locally Accessible Inhibitor Conformations in Drug-Bound HIV-1 Protease

X-RAY crystallography provides an invaluable insight into the tertiary structure and function of proteins. In particular, structures of ligands bound to enzymes reveal the structural properties of such ligands in complex, as well as improving our understanding of the steric mechanisms of active site binding. Furthermore, the emergence of crystal structures for a wide range of proteins has led to an evolution in the pharmaceutical industry, where pharmacologically interesting targets form the basis of "structure guided-drug design" principles. However, one drawback of crystal structures is that they provide only a limited degree of information regarding the dynamical behaviour of proteins. Proteins are mobile entities and alongside their structural properties, a rigorous understanding of their dynamics at timescales ranging from the ns to the ms is essential for a complete understanding of their function. Furthermore, whilst crystal structures provide information into a set of possible ligand conformations and thus interactions that can exist between ligand and protein, they are by no means exhaustive and it is within this context that protein-ligand interactions can benefit from a more dynamical treatment.

In Chapter 2 we described the theory of molecular dynamics and its applicability in studying the dynamics of biological macromolecules. In this chapter we use multiple molecular dynamics simulations to investigate the protein-ligand interactions mediated by wildtype and mutant HIV-1 proteases complexed to the inhibitor saquinavir. Our aim is to complement existing crystal structures of such complexes by exploring the multitude of drug conformations that may be possible, and to subsequently investigate the differences between them.

This chapter is therefore organised subsequently into a 'Background' section in which we provide a more detailed overview to the design of peptidomimetic inhibitors as well as the drug resistant mutations characteristic to saquinavir. This is followed by a 'Methods' section, in which we describe the specific details regarding the implementation of our molecular dynamics simulations and post-simulation anal-

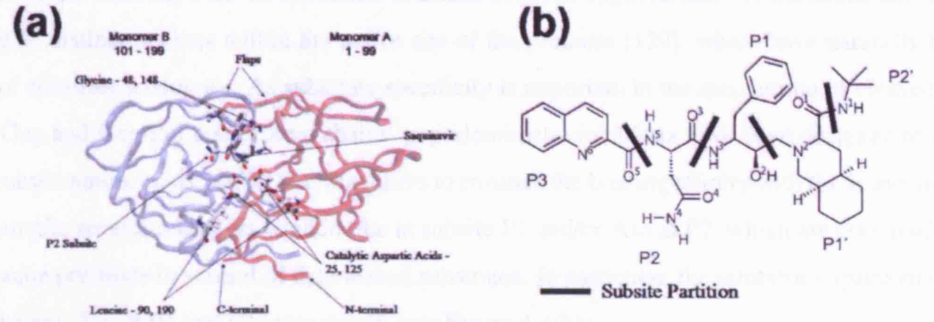


Figure 4.1: (a) Dimeric three dimensional structure of HIV-1 protease complexed with the inhibitor saquinavir (extracted from 1HXB). We label residues of monomer A (red) from 1 to 99 and those of monomer B (blue) from 101 to 199. The positions of residues 48/148 and 90/190 at which the primary saquinavir resistant mutations occur are shown. (b) The chemical structure of saquinavir showing the P1-P3 and P1′-P2′ subsites with polar atoms labelled O^1 - O^5 and N^1 - N^6. The P2 subsite of saquinavir is identical to the end of an asparagine side chain. The exact partitioning of each subsite adopted in this study is also shown (black line).

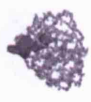
ysis, a 'Results' section in which we report our findings and finally a 'Discussion' section.

4.1 Background

The protease synthesised by the human immunodeficiency virus (HIV) is a C_2-symmetric dimer [113] that encloses a pair of aspartic acid residues (one from each monomer), responsible for proteolytic cleavage [120] in the active site (see Figure 4.1(a)). The active site is bound by a pair of hairpin β-sheets or 'flaps' that allow the lytic regions of the Gag and GagPol polyprotein substrates cleaved by the protease access to the aspartic acid dyad.

Although recent studies have shown a possible allosteric site as a target for the design of novel inhibitors [114], conventional inhibitors of HIV protease have targeted the active site using both symmetric and asymmetric peptidomimetic principles to achieve inhibition. The strategy employed in designing such inhibitors has been to use the peptides that represent the lytic substrate regions as a template from which to design substrate analogues [241]. The cleavable peptide bond is usually replaced by a hydroxyethylene group that can form hydrogen bonds with the dyad; several inhibitors have also been found to utilise a centrally bound water molecule between the drug and the flaps to aid inhibition [166].

Side groups along from the cleavable peptide bond in the N-terminal and C-terminal directions of these substrates are labelled sequentially as subsites P1, P2, P3 and P1′, P2′ and P3′ respectively [128].



Subsites in the naturally cleaved substrates alternate between opposite sides of the substrate backbone, binding to distinct pockets within the active site of the protease [129], which have naturally been the target of substrate analogues. As substrate specificity is important in the recognition of cleaved regions of the Gag and Gag-Pol polyprotein chains, peptidomimetic inhibitors have been designed to preserve some subsite amino acids, whilst altering others to enhance the binding affinity with the active site [241]. For example, several inhibitors contain Phe at subsite P1 and/or Asn at P2, which are commonly found at the same positions in several of the cleaved substrates. In particular, the inhibitor saquinavir contains both Phe and Asn at P1 and P2 respectively (see Figure 4.1(b)).

Unfortunately, inhibitor efficacy is severely limited by the emergence and proliferation of drug resistant mutations characteristic of each inhibitor [173]. In particular, treatment with saquinavir has led to the emergence of primary mutations G48V and L90M, where G48V denotes glycine at residue position 48 mutating to valine on each monomer and L90M similarly denotes mutation from leucine to methionine at position 90. Unlike the G48V mutation, which interacts directly with the inhibitor, the L90M mutation cannot directly influence the drug as residue 90 lies in the α-helix of each monomer that supports the base of the active site (see Figure 4.1), and is therefore buried in the protease.

Structural information is an invaluable starting point for the description of intermolecular ligand-protease behaviour [132]. However, molecular dynamics simulations are also crucial in providing insight [136–138, 142], especially in situations where structural differences between proteases remain small, yet where significantly different clinical behaviour is observed [139]. The L90M mutation is an example of this.

In this chapter we investigate the conformational properties of the P2 subsite of saquinavir in the context of the aforementioned characteristic drug resistant mutations. In previous studies, the P2 subsite of saquinavir has been shown across two 1 ns simulations to adopt different conformations in the wildtype and G48V mutant [242], hydrogen bonding to residue 48, which lies near the tip of the flaps in the wildtype, but rotating away from the flaps in the G48V mutant. Given the crucial role played by the P2 subsite in substrate recognition as well as inhibitor design, it is particularly important to extend such conformational studies in order to explore the multitude of conformations that may be exhibited by it in the active site of both wildtype and mutant proteases, particularly since two distinct conformers are observed in existing x-ray structures.

Previous studies on crambin have shown that single molecular dynamics simulations may not be adequate to sample the extent of the conformational landscape available to a macromolecule [52] due to the limitations of the phase space explored. Instead multiple simulations with slightly varying initial conditions, were found to be necessary to allow the molecule to overcome energetic barriers unscalable in single simulations and to sample more extensive regions of the phase space available to it.

Here, we adopt such a strategy in the study of the multiple conformations of the P2 subsite in order to more fully characterise the conformational landscape available to the drug in the active site of the

protease. This is implemented by performing an ensemble of fifteen molecular dynamics simulations on each of four saquinavir complexes, namely the wildtype, the G48V and L90M single mutants and the G48V/L90M double mutant. Furthermore we investigate how the various conformational minima observed alter both the hydrophilic and hydrophobic interactions between the drug and the protease as well as drug-protease interaction energies and to what extent mutations alter the distribution of the distinct conformations available.

The study reported here is related to the study presented in Chapter 5, in which we report that the effect of coupling to distinct regions of the active site by the P2 subsite leads to differential motion of the inhibitor over longer timescales.

4.2 Methods

4.2.1 Initial Preparation of Models

There are currently only two resolved crystal structures available (1HXB 2.3 Å and 1FB7 2.6 Å resolution) for dimeric HIV-1 protease complexed with saquinavir. The 1HXB structure is of the wildtype protease, whilst 1FB7 is of the G48V/L90M mutant. 1HXB was used as the starting point for all the molecular dynamics simulations. The residues of protease monomers labelled A and B in the crystal structure were numbered 1-99 and 101-199 respectively. The crystal structure contains two resolved rotationally symmetric sets of coordinates for saquinavir. These correspond to the rotational symmetry of the dimer and are thus representations of the two ways in which the drug can be positioned in the active site with respect to each protease monomer. The second set of drug coordinates were extracted and missing hydrogens inserted using the PRODRG tool [243]. Gaussian 98 [90] was used to perform geometric optimisation of the inhibitor at the Hartree Fock level with 6-31G** basis functions. The Restrained Electrostatic Potential (RESP) procedure, which is part of the AMBER 7 package [244], was used to calculate the partial atomic charges (see Appendix B). The forcefield parameters for the inhibitor were completely described by the General Amber Force Field (GAFF) [28]. GAFF has been used before in a comparison between saquinavir and a second generation inhibitor [245]. Mutations on the protease were incorporated using a protocol from the visualisation package VMD [50] which also inserted all missing hydrogens on the protease.

The crystal structure for the G48V/L90M double mutant complexed with saquinavir was also available (1FB7) and was a possible starting point for the double mutant simulation. However, as crystal structures of the single mutant complexes did not exist, and thus employing a mutational protocol was inevitable, the 1FB7 structure was discarded in order for the double mutant to remain consistent with respect to both the single mutants and the wildtype. Instead the mutational protocol was employed to construct the double mutant from the wildtype (1HXB) crystal structure and then compared to the 1FB7

crystal structure (see § 4.3.1). Comparisons of many resolved crystal structures of HIV proteases with significant sequence divergence support the stability of tertiary structure to such mutations [132].

The standard AMBER forcefield for bioorganic systems (ff99) [24] was used to describe the protein parameters. A limitation of the ff99 forcefield is that it over-stabilises α-helical peptide conformations [246]. Recently, a modification of the ff99 forcefield, which improves upon these limitations has been developed [247]. However use of ff99 is not likely to adversely effect the dynamics of HIV-1 protease, given its predominantly β-sheeted structure.

The protonation state of the aspartic acid dyad, which can vary under different conditions (see Kovalskyy *et al.* [156] and references therein), was also considered. Previous molecular simulations on the protease complexed with saquinavir have suggested a monoprotonated dyad with Asp 25 being thermodynamically favoured [242]. However, at physiological pH the catalytic dyad is dianionic and so the proton would have to bind after or upon ligand binding. Therefore, as an equilibrium is likely to exist between monoprotonated and dianionic states, in this study we adopted a model of the dyad in the dianionic state. Studies on the thermodynamically favoured monoprotonated state were also conducted, and are reported in Chapter 6.

The Leap module [248] in the AMBER 7 software package [244] was then used to combine each apo-protease system with the inhibitor, whilst retaining the crystal structure water molecules. Four Cl^- counter-ions were added to electrically neutralise each system, which was then solvated using atomistic TIP3P water [249] in a cubic box with at least 10 Å distance around the complex. The size of each prepared system was 31845, 31860, 31841 and 31856 atoms for the wildtype, G48V, L90M and G48V/L90M systems respectively.

4.2.2 Minimisation and Equilibration Protocols

The molecular dynamics package NAMD2 [34] was used throughout the production simulations as well as for the employment of minimisation and equilibration protocols. Minimisation was conducted using the conjugate gradient and line search algorithms available in NAMD2 for 700 iterations for each system with a force constant of 25 kcal/mol/Å^2 applied to all restrained atoms. This achieved a desired gradient tolerance of between 10 eVA^{-1} to 20 eVA^{-1} in each case. Restrained atoms included all heavy atoms of HIV-1 protease and saquinavir.

The equilibration protocol was adapted from Perryman *et al.* [114] with several important modifications. The long range Coulombic interaction was handled using the particle mesh Ewald summation method (PME) [250]. A non-bonded cutoff distance of 12 Å was used for all simulations. For the equilibration and subsequent production run the SHAKE algorithm [38] was employed on all atoms covalently bonded to a hydrogen atom, allowing for an integration timestep of 2 fs. Each system was gently annealed from 50K to 100K over a period of 10 ps. This was followed by further annealing to 300K over a period of 20 ps. The systems were then maintained at a temperature of 300K using

a Langevin thermostat with a coupling coefficient of 5 /ps for the rest of the equilibration and for all subsequent production runs. The systems were equilibrated for 200 ps whilst maintaining the force constants on the restrained atoms to allow for thorough solvation of the complex and to prevent premature flap collapse [251].

The L90M and G48V/L90M systems were then equilibrated for 50 ps with the force constraints of all atoms within a 5 Å radius of the L90M mutation and their respective residues set to 0. This was to allow optimal re-orientation of the substituted methionine which would normally be inaccessible due to the burying of the residue beneath the surface of the protease. For this reason it was not necessary to allow the G48V mutation any special re-orientation time as it exists in the flap, is completely exposed to the surface and thus able to access conformations which are energetically favourable in the natural course of the equilibration. The atoms within the 5 Å radius of L90M were then kept unrestrained for all further simulations.

The next stage entailed a gradual force reduction on the restrained atoms. The force constant was reduced sequentially to 20, 15, 10 and 5 kcal/mol/Å^2, equilibrated for 50 ps at each value. This was followed by a completely unrestrained isothermal equilibration for 300 ps. This concluded the equilibration in the canonical (NVT) ensemble which lasted a total of 730 ps for wildtype and G48V systems and 780 ps for the L90M and G48V/L90M systems.

4.2.3 Production Ensembles

The initial production phase for each simulation was implemented in the NVT ensemble and continued directly from the last step of the equilibration phase for a duration of 500 ps. The entire equilibration process and subsequent production was repeated from the same initial conditions of minimised energy to construct an ensemble of fifteen production trajectories for each protease system. Each simulation within an ensemble varied only in the initial velocities attributed to the constituent atoms, which in each case were randomised in a way that reproduced the Maxwell-Boltzmann distribution for a given temperature.

Several of the simulations within each ensemble were extended for a further period of 2 ns to evaluate the stability of the P2 subsite conformations of saquinavir (see Table 4.1). These simulations employed an isothermal-isobaric ensemble (NPT) using a Berendsen barostat [49] with a target pressure of 1 bar and a pressure coupling constant of 0.1 ps. Coordinate trajectories were recorded every 1 ps throughout all equilibration and production runs.

In total, approximately 100 ns of simulation was achieved from the equilibration and production ensembles and selected 2 ns extensions. The simulations were performed under conditions of optimal computational efficiency using NAMD2 [34], with a wall-clock rate of approximately 8 hours/ns, using 30 processors on a 512 processor SGI Altix at CSAR, University of Manchester, UK, 32 processors (1 node) at the UK national HPCx facility, Daresbury, and 32 processors on the TeraGrid cluster at NCSA.

These simulations also made use of 32 processors of the Leeds and Oxford compute nodes of the UK National Grid Service.

4.2.4 Post-Production Analysis

The simulations were analysed using a range of methods. The root mean squared fluctuation (RMSF) of the protease backbone across all simulations conducted in the wildtype and mutant ensembles were calculated over the 500 ps of the initial NVT production phase. The average structure of the backbone across this time was taken as the reference structure. The initial conformation of the P2 subsite adopted in each run was analysed first by visual inspection using VMD and secondly by calculating the dihedral angle and hydrogen bond distances characterising each conformation. One occurrence of each conformation observed in each of the ensembles was selected as a representative run for that conformation in that particular ensemble. These runs, which were extended for a further 2 ns in the NPT ensemble as described above, were subsequently analysed. The effects of restrained solvation on the protease caused cavities to form in the solvent around the edge of the simulation box in the NVT ensemble. Equilibration in the NPT ensemble was thus achieved by reduction of the box volume to a new equilibrium value within 500 ps for each run. Therefore, to allow for such re-equilibration in the new ensemble, subsequent analysis was only conducted on the last 1 ns of each representative extended run.

Gas-phase interaction energies $\langle \Delta V \rangle$ were calculated as the sum of the van der Waals, $\langle \Delta V \rangle_{vdW}$ and electrostatic, $\langle \Delta V \rangle_{ele}$ contributions to the interaction between the drug and the protease.

$$\langle \Delta V \rangle = \langle \Delta V \rangle_{vdW} + \langle \Delta V \rangle_{ele} \tag{4.1}$$

$$\langle \Delta V \rangle = \frac{1}{N} \sum_{i=1}^{N} V_i^{com} - (V_i^{pro} + V_i^{lig}) \tag{4.2}$$

The difference between the interaction potential for the complex and the sum of the separate protease and drug potential energies then gave the interaction energy. The interaction energy was averaged over 100 equally distributed snapshots (N = 100) over the selected 1 ns of each trajectory analysed (1 each 10 ps). The molecular mechanics component of the MMPBSA module in the AMBER 9 software package [53] was used to implement the analysis.

Hydrogen bond and hydrophobic contact analysis was implemented using 'tcl' scripts developed for use with VMD. Donor and acceptor atoms for the drug and the subsets of protease residues considered were explicitly defined as all polar nitrogen and oxygen atoms. The criteria for the existence of an instantaneous hydrogen bond were a donor-acceptor distance $\leq$ 3.5 Å and a donor-hydrogen-acceptor angle $\geq$ 150°. Instantaneous hydrogen bonds were calculated each 1 ps and the mean number of hydrogen bonds between two sets of atoms considered was calculated as the time average of all instantaneous hydrogen bonds. The frequency of occurrence of singular hydrogen bonds was calculated in

the same way. The criterion for the existence of instantaneous hydrophobic contacts was an atom-atom distance $\leq$ 3.5 Å. The mean number of hydrophobic contacts was calculated using the same method as that for the hydrogen bond analysis. Hydrophobic atoms in the protease were implicitly defined by VMD, whilst in the drug, they were explicitly defined as all non-carbonyl carbon atoms and all non-polar hydrogen atoms. Finally drug subsite analysis was implemented by partitioning the atoms of the drug belonging to each subsite as shown in Figure 4.1(b).

The positional dynamic cross correlation map (DCCM) of the amino acid backbone was calculated for wildtype HIV-1 protease bound to saquinavir. The last 1 ns of the R2 trajectory in NPT of the wildtype ensemble was used for the analysis. A methodology similar to that of Zoete *et al.* [132] was adopted in calculating the DCCM. The trajectory was partitioned into 20 blocks of 50 ps. Each block therefore had 50 snapshots from which to calculate cross correlations. The positional cross-correlation coefficients between the geometric centres of the backbone of each amino acid residue were calculated for each block using the method discussed in § 2.5.1. The cross correlation matrices of all 50 ps blocks were then averaged to construct a single DCCM (see § 4.3.1).

4.3 Results

4.3.1 Overall Structural Characteristics

Characteristics of saquinavir-bound crystal structures

It is important to provide a preliminary description of the similarities and differences of the key ligand-protease interactions in the crystal structures of saquinavir bound to HIV-1 protease. Figure 4.2(a) shows the structure of the active site of the protease formed by the S3-S2′ subsites in response to binding with saquinavir in the 1HXB crystal structure. Figure 4.2(b) shows the active site as viewed from the opposite lateral entrance. The shape of the subsites is depicted using a solvent accessible surface area plot with a probe radius of 1.4 Å. Figures 4.2(c) and 4.2(d) show the structure of the P3-P2′ subsites of the drug with respect to these protease subsites. Although, the exact shape of subsites S3-S2′ are to some degree dependent on the structure of the inhibitor subsites, the overall structure is largely independent. Thus, owing to the rotational symmetry of the protease, the S1 (red) and S2 (blue) subsites are similar to the S1′ (orange) and S2′ (yellow) subsites respectively.

Furthermore, we have defined the protease subsites as those formed by all residues within 4 Å of the corresponding inhibitor subsite. Therefore, some subsites partially overlap due to the fact that each inhibitor subsite may interact with more than one specifically designated S-subsite. Figure 4.3 shows the decomposed structure of each protease and inhibitor subsite. The S3 subsite (green), which accommodates the P3 inhibitor subsite, is principally composed of G148, G149 in the flaps and the hydrophobic P81 residue as well as R8, G127, A128, D129 and D130, which make marginal contact with P3 (see Fig-

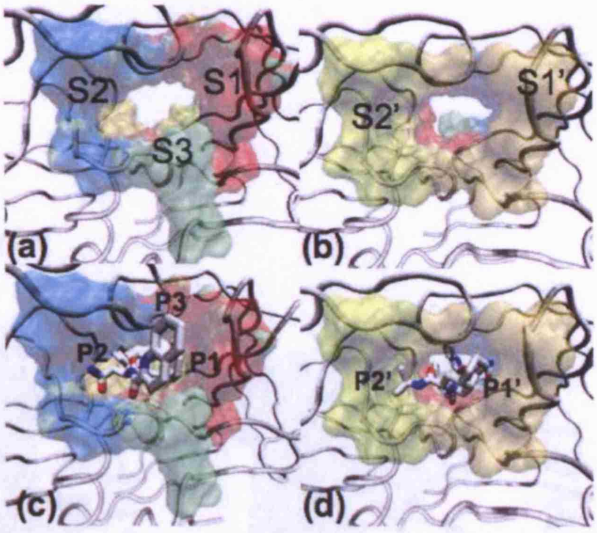


Figure 4.2: Overall structure of protease and saquinavir subsites in the 1HXB crystal structure. (a) Structure of the active site composed of the S3-S2′ subsites as depicted using a solvent accessible surface area plot with a probe radius of 1.4 Å. (b) View of the active site from the opposite lateral entrance. The drug has been removed in both cases. (c) and (d) show the structure and position of the P3-P2′ drug subsites with respect to the corresponding protease subsites.

ure 4.3(a)). The S2 subsite (blue) is amphiphilic, being principally composed of hydrophobic residues I50, I184, I147, V132 and A128 and hydrophilic residues D129 and D130. The P2 subsite is identical to asparagine and is in close proximity to D129, D130 and the carbonyl carbon of G148 (see Figure 4.3(b)). Interestingly, due to the orientation of the P2 subsite, it is not within 4 Å of I184 in the 1HXB structure. The S1 subsite (red) is principally composed of hydrophobic residues I150, P81, V82, I84 and L23 which accommodate the phenyl group of the P1 subsite of saquinavir. The hydroxyethylene moiety of the P1 subsite also interacts hydrophilically with the catalytic dyad D25/D125 as well as being in proximity to G127 and A128 (see Figure 4.3(c)). The S2′ subsite (yellow) is composed of the same residues as the S2 subsite but those for the alternate monomer of the protease. The P2′ subsite additionally is within 4 Å proximity to the I84 residue and is orientated towards the hydrophobic core of residues (I150, I47, V32, I84 and A28) that form part of the P2′ subsite (see Figure 4.3(d)). The S1′ subsite (orange) is also composed of the same residues as S1 but for the alternate monomer. The hydrophobic P1′ subsite is orientated towards the pocket composed of similarly hydrophobic residues I50, P181, V182, I184, L123 and additionally T180 (see Figure 4.3(e)). Finally, a tetrahedrally coordinated water molecule is bound between the flaps of the protease and the drug, specifically to the backbone nitrogen atoms of residues I50 and I150 and the O^1 and O^3 atoms of drug subsites P1′ and P2 respectively (see Figure 4.3(f)).

Figure 4.4(a) shows the superposition of saquinavir in the active site of HIV-1 protease from both the

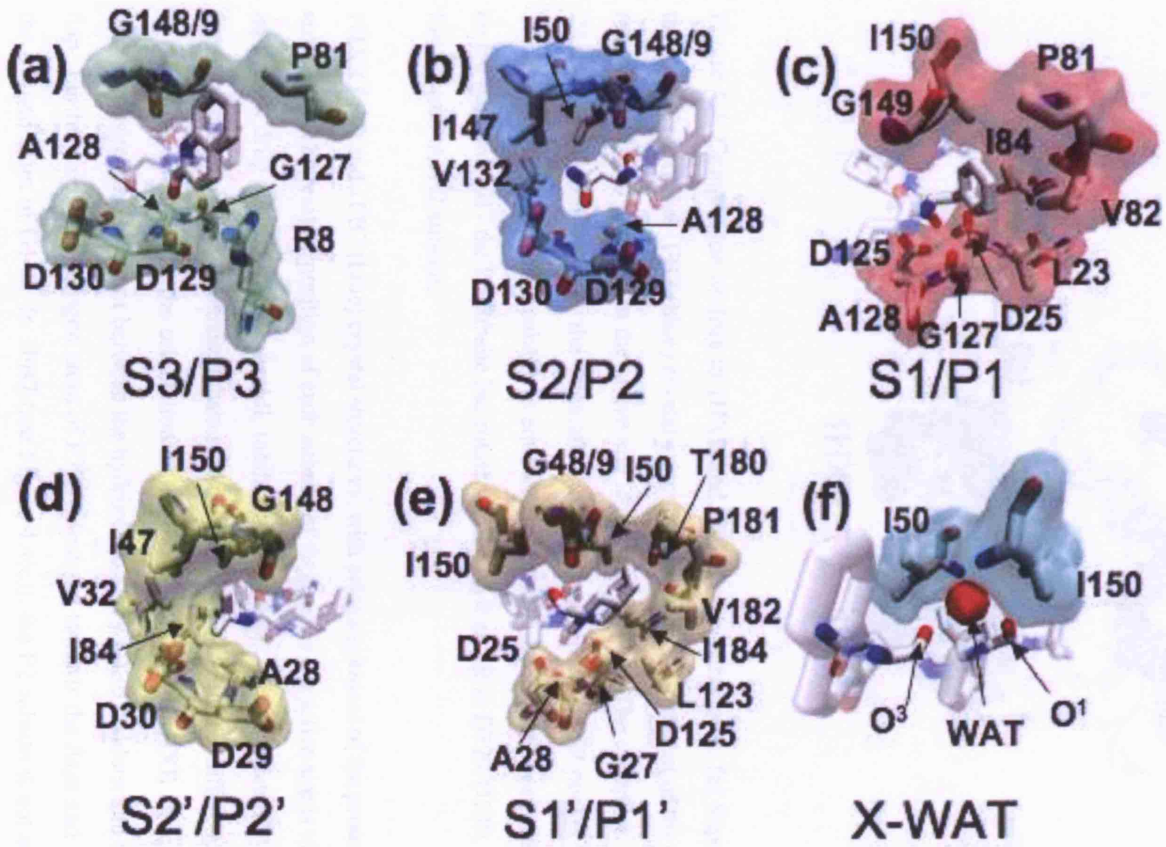


Figure 4.3: Decomposed structure of protease and saquinavir subsites in the 1HXB crystal structure. The amino acid residues that compose each of the protease subsites as well as the corresponding molecular structure of the drug subsites which interact with them are shown specifically for the (a) S3/P3, (b) S2/P2, (c) S1/P1, (d) S2′/P2′ and (e) S1′/P1′ subsites. (f) shows the tetrahedrally coordinated hydrogen bond interaction of the crystallographic water molecule in between the flaps of the protease and the inhibitor.

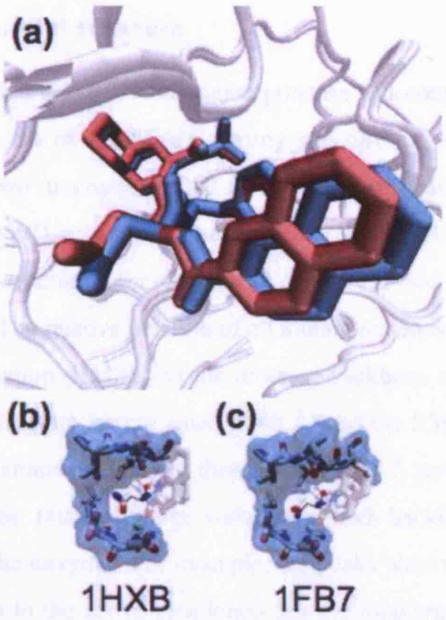
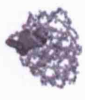


Figure 4.4: Comparison of drug in 1HXB and 1FB7 crystal structures. (a) Superposition of the drug in the 1HXB (red) and 1FB7 (blue) crystal structures, after prior alignment of the protease backbone. The relative structural position in the active site is almost identical. The exception is the orientation of the P2 subsite. (b) and (c) show the orientation of the P2 subsite in the S2 pocket in structures 1HXB and 1FB7 respectively. The P2 subsite is equadistant between hydrophilic residues D129/D130 and G148 on the flap. In 1FB7 the P2 subsite has rotated, bringing it closer to D129/D130, as well being orientated further into the S2 subsite.

1HXB (red) and 1FB7 (blue) crystal structures, with prior alignment of the protease backbone in the two structures. The relative position of each subsite of the drug in the active site is virtually identical, giving an overall drug RMSD of 0.71 Å. All, inhibitor subsites are accommodated by the same residues that form the corresponding protease subsites. Indeed, the only significant difference in drug conformation in these two structures is the orientation of the P2 subsite. In the 1HXB structure (see Figure 4.4(b)) the P2 subsite is equadistant between the hydrophilic D129/D130 residues and the G148 residue on the flap. Furthermore, the nitrogen atom of P2 is orientated towards the flaps and is in close proximity to the oxygen atom of G148. In 1FB7 (see Figure 4.4(c)), the P2 subsite is not only closer to D129 and D130, but also rotated with respect to 1HXB so that the nitrogen atom is orientated further into the S2 subsite.



Comparison of model with crystal structure

The structure of the 1HXB-derived G48V/L90M mutant protease was compared with the 1FB7 crystal structure in order to validate the use of 1HXB as a starting structure for wildtype and all three mutant systems. The backbones of the two structures, 1HXB-G48V-L90M (red) and 1FB7 (blue), were aligned prior to measurement of the RMSD across protease residues (see Figure 4.5(a)). Visually, the most notable differences in backbone structure is for residues 37-41 and 137-141 corresponding to the flexible loop regions of each monomer. The relative position of all mutant residues is almost identical.

The root mean squared deviation (RMSD) of the protease backbone of the 1HXB-derived double mutant compared to the 1FB7 structure is very small (0.48 Å) and the RMSD values for the backbone and side chain atoms of mutated amino acids is less than 0.5 Å and 1.5 Å respectively (see Figure 4.5(b)). Regions of the protease that show relatively large variation in both backbone and side-chain RMSDs correspond to flexible parts of the enzyme. For example, the peaks observable at residues 39/139 and 150 in Figure 4.5(b) correspond to the above-mentioned flexible loop region leading into the flaps as well as the flap tips respectively, both of which have been shown to exhibit structural differences in previous studies [132].

Dynamic Cross-Correlation Map of Wildtype HIV-1 Protease

In order to validate the accuracy of the molecular simulation protocol in reproducing the global structural properties of the protease, we determined the dynamical cross correlations between the backbone atoms of each amino acid residue across the protease (see Figure 4.6). The principal secondary structure characteristics were well reproduced. Strong correlations were observed for the β-sheets forming each monomer such as the fulcrum (residues 10-20), flaps (residues 43-58) and cantilever (residues 60 - 77) characterised by perpendicular extensions to the diagonal. Furthermore, the characteristic strong correlations between residues 10-20 with 60-70 and residues 70-90 with 25-35 were also exhibited. The correlations were in good agreement with cross-correlation maps derived from the comparison of crystal structures of HIV-1 protease (see § 3.4.3 and Zoete *et al.* [132]).

4.3.2 Multiple Conformations of the P2 Subsite

Using 1 ns molecular dynamics, previous authors have reported that the amide end of the P2 subsite of the drug, which is identical to the asparagine side chain of the drug, forms a strong hydrogen bond with the peptide carbonyl oxygen of residue 148 of the wildtype protease [242]. These studies also showed that in the G48V mutant the P2 subsite rotates away from this carbonyl oxygen, breaking the hydrogen bond and thus decreasing the interaction of saquinavir with the flaps.

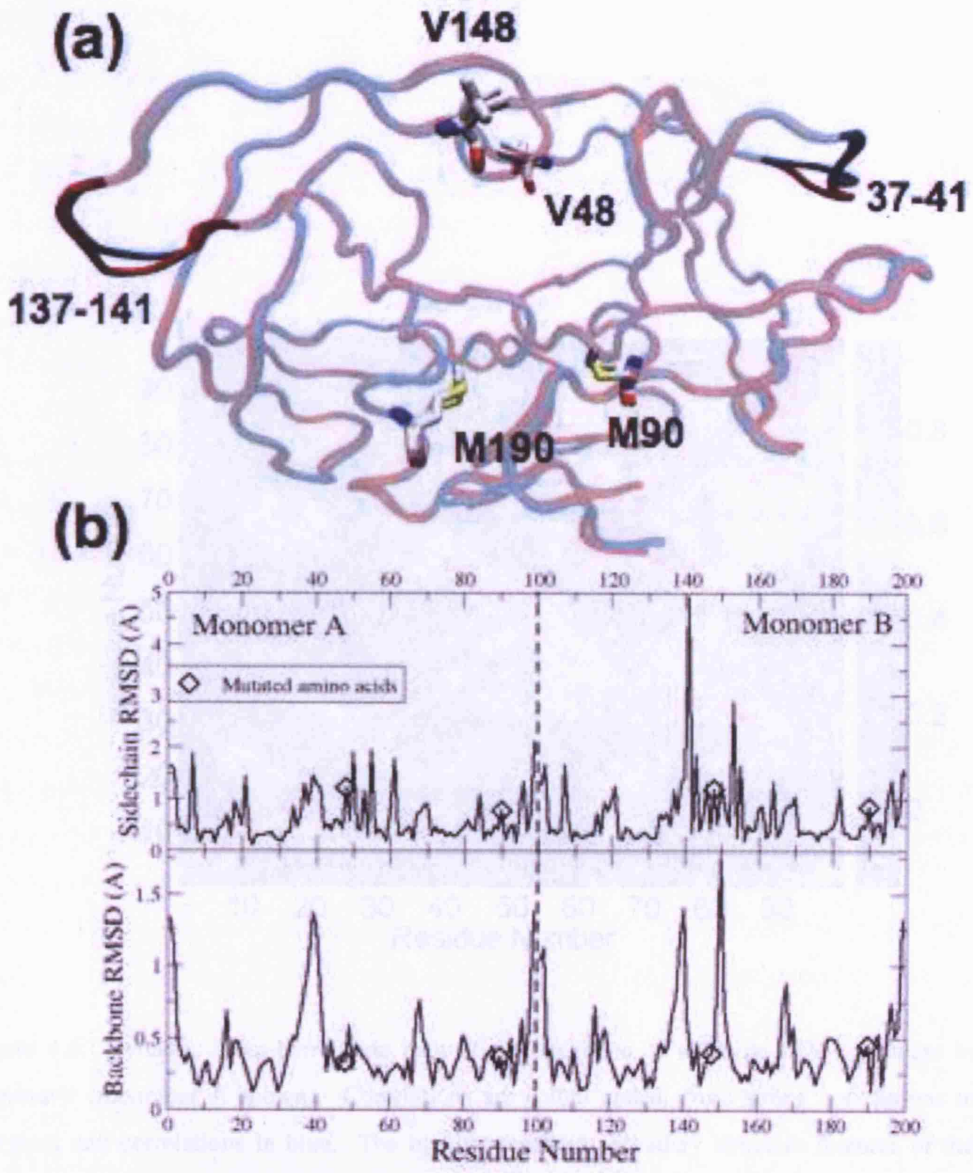
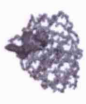


Figure 4.5: Comparative validation of the 1HXB-derived G48V/L90M mutant and the 1FB7 mutant crystal structure. (a) Schematic diagram of aligned protease backbones in 1HXB-derived mutant (red) and 1FB7 (blue). The flexible loop regions 37-41 and 137-141 are highlighted as well as the structure of the M90/M190 and V48/V148 mutant residues. (b) RMSD of backbone (bottom) atoms and non-hydrogen side-chain atoms including backbone (top) against residue number for the 1HXB-G48V-L90M starting structure relative to 1FB7. Backbone RMSDs of mutated amino acids are less than 0.5 Å, showing strong equivalence of implemented mutations to the crystal structure of 1FB7. Side-chain RMSDs of mutated amino acids are less than 1.5 Å. All large deviations ($>$ 2 Å) correspond to regions of the protease that are naturally flexible, such as the flaps and the flexible loop regions.



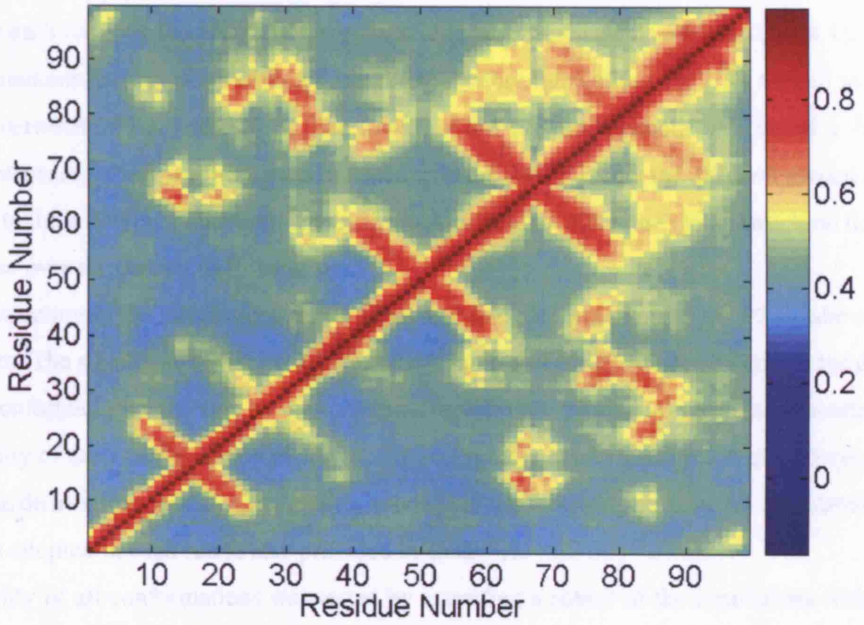
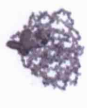


Figure 4.6: Dynamic cross-correlation map of the backbone of wildtype HIV-1 protease bound to saquinavir (monomer A shown). Correlations are colour coded, from strong correlations in red to marginal anti-correlations in blue. The highly-correlated secondary structure features of the amino acid backbone are all present and agree well with cross-correlation maps derived from analysis of multiple crystal structures (see § 3.4.3 and Zoete *et al.* [132]). These include β-sheets characterised by perpendicular extensions to the diagonal as well as broadening of the diagonal representing α-helices (residues 85-95). The characteristic strong correlations between residues 10-20 with 60-70 and residues 70-90 with 25-35 are also exhibited.



Ensemble Characteristics

We conducted an ensemble of 15 simulations (each of 1.3 ns including equilibration) for every protease/saquinavir complex. Each run in the ensemble varied only by the initial velocities assigned to the atoms according to a randomised Maxwell-Boltzmann distribution, fitting the initial temperature. The last phase of the equilibration and the 500 ps production phase was conducted in the canonical ensemble (NVT) and in order to validate the establishment of a thermodynamically equilibrated system, the temperature fluctuations of the system and the root mean squared fluctuations (RMSF) of the protein backbone for each run over the 500 ps of the production phase were calculated (see Table 4.1).

Every simulation, except one, exhibited small temperature fluctuations (< 2 K) as well as a mean temperature between 299 K and 300 K. The RMSF of the protein backbone was below 1 Å for all systems. Furthermore, there was no significant change in the RMSF across the different mutant systems as compared to the wildtype, indicating that these mutations do not affect the global structural flexibility of the protease over the picosecond timescale.

Visual inspection of the simulations showed that multiple conformations of the P2 subsite can exist in each system. The significant initial flexibility and rotatability of the P2 subsite, leading to the diversity of different conformations exhibited by it, prompted a further investigation into both the characterisation and the stability of each conformation observed. Although these conformations are characterised using dihedral angle distributions and distinct hydrogen bonds in the following section, for completeness, the conformation adopted in each run is also provided in Table 4.1.

The stability of all conformations was tested by extending a subset of the simulations within each ensemble. Due to computational restrictions coupled with the large simulation ensemble size and the 4-fold multiplicity of the ensembles resulting from the wildtype and the three mutants studied here, it was not feasible to extend the simulations for the entire ensemble size. Instead, one representative occurrence of each conformation from every ensemble was selected and extended by a further 2 ns (see Table 4.1). Furthermore, as the conformation C_α was always a precursor of C_β and C_γ and was not characterised by any significant hydrogen bonds (see § 4.3.2), we also extended simulations for all occurrences of this conformation in every ensemble again for a duration of 2 ns.

All extended simulations were implemented in the NPT thermodynamic ensemble. In order to allow for re-equilibration of the systems in the new ensemble, which always occurred within the first 500 ps, subsequent analyses of the systems were carried out for the last 1 ns of the extended trajectories.

Conformational Characterisation and Stability Analysis

The conformations were characterised by the association of the P2 subsite with distinct residues within the active site of the protease as well as by two specific P2 subsite dihedral angles, C^{25}-C^{26}-C^{27}-C^{28} and C^{26}-C^{27}-C^{28}-N^4, termed χ_1 and χ_2 respectively (see Figure 4.7).

	Wildtype			G48V			L90M			G48V/L90M		
Run	$\langle T \rangle$ (K)	RMSF (Å)	C_{P2}	$\langle T \rangle$ (K)	RMSF (Å)	C_{P2}	$\langle T \rangle$ (K)	RMSF (Å)	C_{P2}	$\langle T \rangle$ (K)	RMSF (Å)	C_{P2}
R0	299.18 (1.66)	0.60	C_β	299.38 (1.68)	0.62	C^*_δ	299.27 (1.68)	0.65	C^*_δ	299.23 (1.66)	0.93	C^*_ε
R1	299.24 (1.66)	0.63	C_β	299.35 (1.65)	0.58	C_δ	299.34 (1.63)	0.57	C_δ	299.30 (1.65)	0.57	C^*_β
R2	299.36 (1.65)	0.61	C^*_β	299.30 (1.63)	0.57	C_δ	299.26 (1.66)	0.97	C^*_β	299.32 (1.66)	0.57	C^*_γ
R3	299.48 (1.63)	0.60	C_β	299.29 (1.66)	0.62	C_β	299.27 (1.66)	0.65	C^*_α	299.33 (1.67)	0.60	C^*_δ
R4	299.31 (1.66)	0.57	C^{**}_β	299.33 (1.67)	0.55	C^*_ε	299.25 (1.63)	0.57	C_α	299.29 (1.69)	0.78	C_γ
R5	299.21 (1.66)	0.63	C^*_δ	299.24 (1.64)	0.62	C^*_β	299.31 (1.67)	0.54	C_δ	299.35 (1.63)	0.62	C^{**}_ε
R6	299.41 (1.62)	0.60	C^*_ε	299.24 (1.63)	0.55	C_β	299.22 (1.72)	0.88	C_α	299.32 (1.66)	0.55	C_δ
R7	299.37 (1.72)	0.55	C_β	299.33 (1.66)	0.61	C_β	299.22 (1.64)	0.57	C^{**}_β	299.28 (1.64)	0.57	C_ε
R8	299.24 (1.66)	0.64	C^*_α	299.26 (1.60)	0.62	C^{**}_ε	299.20 (1.66)	0.59	C_α	299.36 (1.65)	0.66	C_δ
R9	299.25 (1.65)	0.58	C_α	299.36 (1.67)	0.58	C_δ	299.24 (5.35)	0.56	C_δ	299.32 (1.66)	0.62	C_δ
R10	299.34 (1.68)	0.55	C_β	299.22 (1.64)	0.95	C_δ	299.22 (1.68)	0.58	C_δ	299.25 (1.62)	0.54	C^*_α
R11	299.21 (1.68)	0.79	C_β	299.35 (1.63)	0.68	C_β	299.28 (1.67)	0.56	C_β	299.25 (1.69)	0.55	C_δ
R12	299.27 (1.67)	0.58	C_β	299.39 (1.66)	0.61	C^{**}_δ	299.31 (1.68)	0.56	C_δ	299.30 (1.63)	0.59	C_α
R13	299.31 (1.64)	0.59	C_β	299.28 (1.71)	0.58	C^*_γ	299.27 (1.63)	0.57	C_δ	299.27 (1.64)	0.83	C_δ
R14	299.22 (1.63)	0.58	C_β	299.31 (1.68)	0.55	C_δ	299.09 (1.66)	0.58	C^*_γ	299.31 (1.64)	0.54	C^{**}_β

$\langle T \rangle$ = Mean temperature (values in parentheses denote fluctuations), RMSF = Protein backbone RMSF, C_{P2} = Adopted conformation of P2 subsite

* Denotes selected conformation-representative runs extended for a further 2 ns in the NPT ensemble and for subsequent analysis

** Denotes extended runs initially adopting C_α, but for which transitions were observed away from C_α post-equilibration

Table 4.1: Characteristics of protease-saquinavir ensemble simulations

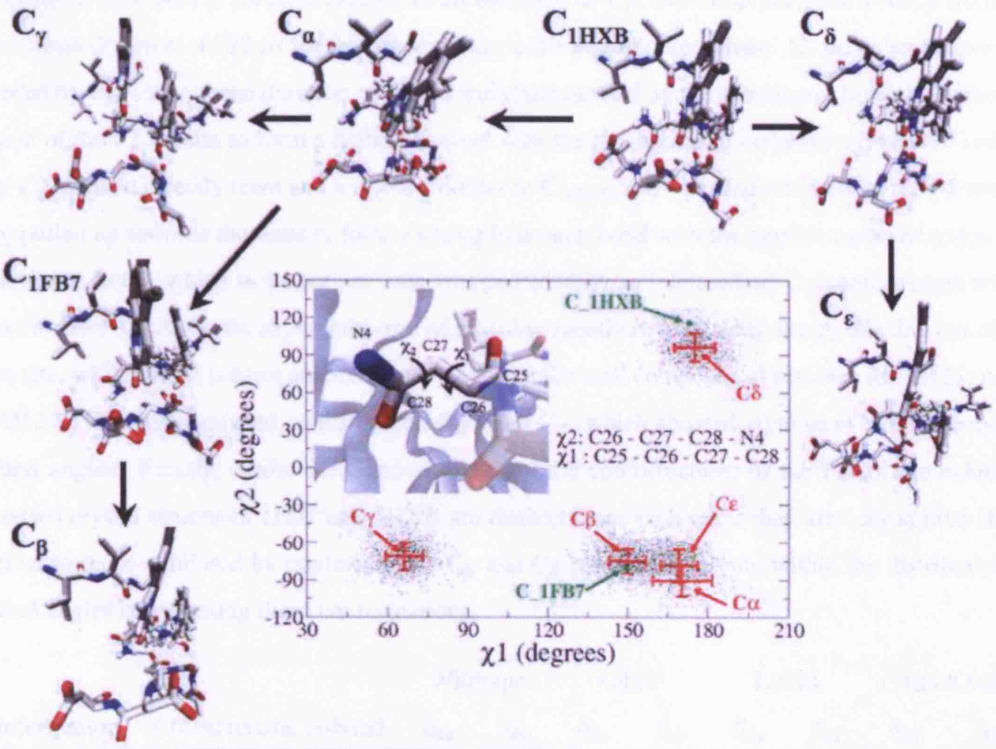


Figure 4.7: Multiple conformations of the P2 subsite of saquinavir. C_{1HXB} and C_{1FB7} are the existing crystal structure conformers. C_α - C_ε were taken from the last 1 ns averaged structures of the G48V/L90M ensemble of simulations. The pathway of evolution of all structures from the starting structure C_{1HXB} is shown. All stable conformations exhibited during the simulations exhibit distinct χ_1 and χ_2 dihedral angle distributions along the P2 subsite, whilst conformations C_β and C_δ are similar to C_{1HXB} and C_{1FB7} respectively.

The initial crystallographic conformation, C_{1HXB}, is orientated with the hydrogens of the P2 subsite pointing towards the flaps. In C_α, the P2 subsite had rotated along the χ_2 dihedral so that the hydrogens pointed downwards, away from the flaps. We identify this as the conformation associated with the G48V mutation reported by previous authors [242]. Conformation C_β was characterised not only by the rotation of the P2 subsite along the χ_2 dihedral but by further subsequent rotation along the χ_1 dihedral, forming a strong hydrogen bond with the oxygen atom OD2 belonging to the catalytic aspartic acid side chain of monomer B (residue D125). In all instances of C_β, evolution occurred directly from C_α with a mean duration of 322 ps for the intermediate in the equilibration phase. C_γ was also exclusively preceded by C_α with a mean duration of 237 ps and characterised by the subsequent inward χ_1 dihedral rotation of the P2 subsite to form a hydrogen bond with the neighbouring carbonyl oxygen (O^3) of the drug. C_δ evolved directly from and was very similar to C_{1HXB}, and was characterised by the P2 subsite being pulled up towards the flaps to form a strong hydrogen bond with the peptide carbonyl oxygen of residue 148. In C_ε, which in all but one case was preceded by an intermediate C_δ conformation with a mean duration of 210 ps, the asparagine arm of P2 subsequently rotated along the χ_2 dihedral out of the active site, while the P2 subsite anchored into a hydrophilic well composed of residues R8, G127, A128 and D129. This distinguished conformation C_α from C_ε, which showed significant overlap between dihedral angles. Finally, whilst the dihedral angles of the conformations of the P2 subsite exhibited in the two crystal structures 1FB7 and 1HXB are distinct from each other they are very similar if not identical to those exhibited by conformations C_β and C_δ respectively, lying within the distribution of dihedral angles representing these conformations.

		Wildtype		G48V		L90M		G48V/L90M	
Conformation	Characteristic H-bond	δ_{da}	f_{hb}	δ_{da}	f_{hb}	δ_{da}	f_{hb}	δ_{da}	f_{hb}
C_α	$N4_{SAQ}$-$OD2_{D130}$	4.52	0.22	na	na	6.46	0.05	5.96	0.00
C_β	$N4_{SAQ}$-$OD2_{D125}$	2.87	0.97	2.84	0.96	2.86	0.94	2.86	0.94
C_γ	$N4_{SAQ}$-$O3_{SAQ}$	na	na	2.99	0.29	2.96	0.31	2.92	0.27
C_δ	$N4_{SAQ}$-$O_{G/V148}$	2.88	0.89	2.91	0.96	3.00	0.88	2.94	0.91
C_ε	$N4_{SAQ}$-O_{G127}	2.88	0.90	2.93	0.64	na	na	2.88	0.80
	$N4_{SAQ}$-N_{D129}	3.50	0.30	3.37	0.35	na	na	3.44	0.37
	$N4_{SAQ}$-$OD1/2_{D129}$	2.90	0.53	2.88	0.60	na	na	2.86	0.75

δ_{da}: Mean donor-acceptor distance (Å), f_{hb}: Frequency of occurrence of hydrogen bond

H-bond criteria: donor-acceptor distance $\leq$ 3.5 Å, donor-hydrogen-acceptor angle $\geq$ 150°

na: Not applicable as conformation did not occur

Table 4.2: Characteristic hydrogen bonds of distinct P2 subsite conformations.

In order to understand the differences in the hydrogen bond arrangements adopted by these confor-

mations, the frequency of occurrence of each ligand-protease and ligand-ligand hydrogen bond formed by the N^4 atom of the P2 subsite was analysed over the last 1 ns of the trajectories (see Table 4.2). The criteria of a maximum donor-acceptor distance of 3.5 Å and a minimum donor-hydrogen-acceptor angle of 150° was used. Mutually exclusive hydrogen bonds between the N^4 atom of the P2 subsite and various polar atoms of either the protease or the inhibitor were characteristic of all conformations, except for C_α, which did not exhibit significant H-bonding between the N^4 atom of the P2 subsite and the active site with a frequency less than 0.25 in all cases. Furthermore, whilst conformations C_β - C_δ were largely defined by singular strong hydrogen bonds associated with the N^4 atom, C_ε formed several hydrogen bonds due to its location within the hydrophilic well described above. Nonetheless the bond with the carbonyl oxygen of G127 could be distinguished as principally characteristic. The significantly reduced hydrogen bond frequency of C_γ, compared to C_β, C_δ or C_ε, is explained by the angular constraints imposed by the P2 subsite when rotating back on itself towards the O^3 atom of the inhibitor. This reduced the frequency with which the stringent donor-hydrogen-acceptor angle criteria were met, even though a small donor-acceptor distance existed between the N^4 and O^3 atoms.

Figure 4.8 shows the donor-acceptor distance for the main hydrogen bonds characterising the P2 subsite conformations C_β - C_ε. These conformations displayed very tight hydrogen bonding with mean donor-acceptor distances all below 3.5 Å. Whilst the dihedral angles exhibited by C_{1FB7} and C_{1HXB} were similar to C_β and C_δ respectively, there were differences in the hydrogen bonds formed by these conformations in the crystal structures as compared to those exhibited by C_β and C_δ. In C_{1FB7}, the P2 subsite had rotated away from the flaps similar to C_β. However, the interatomic distance between atom N^4 of the drug and OD2 of amino acid D125 was 5.65 Å, too large for a hydrogen bond to exist between the two atoms. Therefore, adoption of conformation C_β from C_α must occur via C_{1FB7} with the additional requirement that such a hydrogen bond is formed. In C_{1HXB}, which exhibited a similar P2 subsite orientation with respect to the flaps as compared to C_δ, the distance of the N^4 atom from the carbonyl oxygen O of the backbone of G148 was 4.00 Å. This is very close to the donor-acceptor distances which resulted from a tightly formed hydrogen bond in C_δ and, using a slightly less stringent donor-acceptor distance criterion, would classify C_{1HXB} and C_δ as identical.

For conformations C_β - C_ε, every conformation persisted for the extended 2 ns timescale, exhibiting a significant frequency of occurrence of characteristic H-bonds (see Table 4.2) as well as small mean donor-acceptor distances (< 3.5 Å). Given this, it is a plausible assumption that all occurrences of conformations C_β - C_ε and not just those for which the simulation was extended would persist on this timescale.

The occupation frequency of each conformation in all systems at the end of the aforementioned extended 2 ns simulations is shown in Figure 4.9. From the initial crystallographic conformation C_{1HXB}, the drug in the wildtype moved into conformation C_β more frequently than any other conformation, whilst in the G48V single mutant, the preferred conformation was more equally distributed between C_β

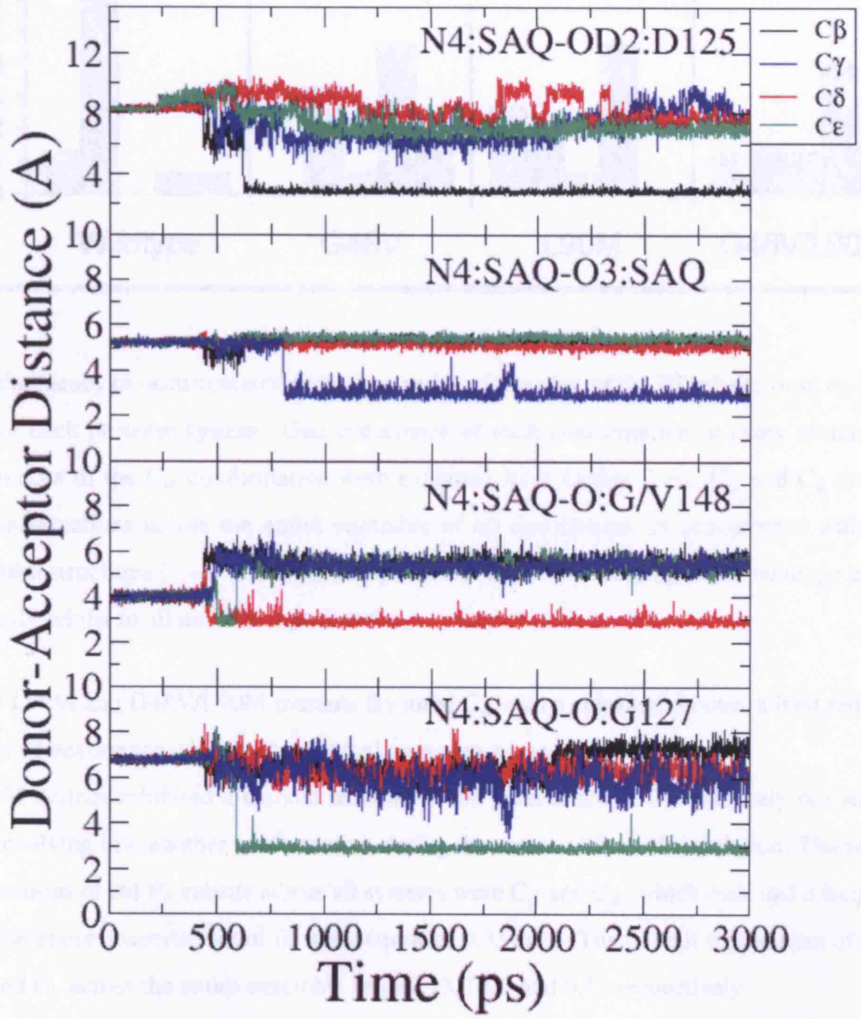


Figure 4.8: Time evolution of donor-acceptor distances for the characteristic hydrogen bonds defining conformations C_β - C_ε of the P2 subsite (for clarity only the evolution of one instance of each conformation is shown). These hydrogen bonds are mutually exclusive and stable over a 2 ns timescale. Adoption of conformation C_ε is preceded by adoption of conformation C_δ for a mean duration of 210 ps in the equilibration phase. C_α forms no significant hydrogen bonds and is not shown here.

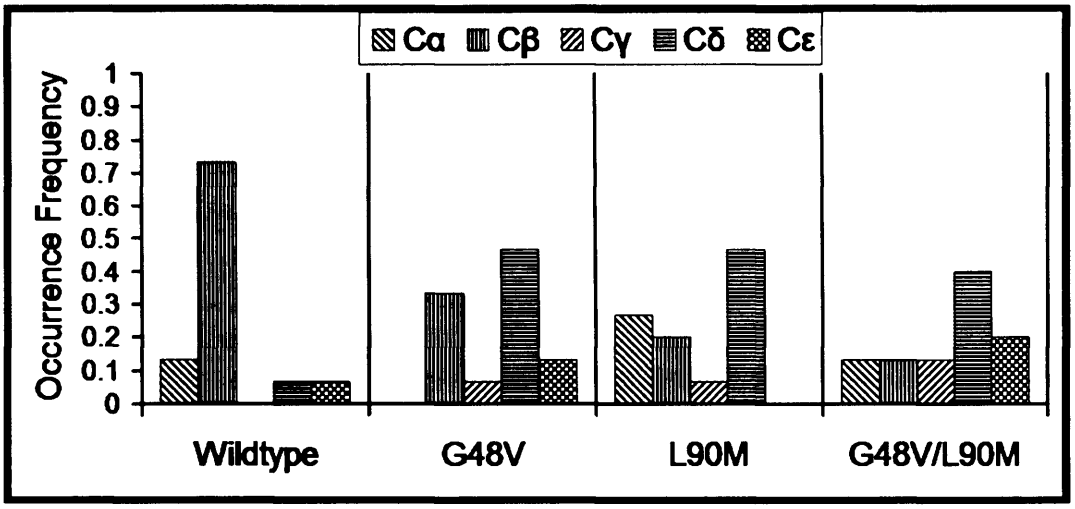


Figure 4.9: Frequency of occurrence of each observed conformation of the P2 subsite from an ensemble size of 15 for each protease system. One occurrence of each conformation in every system as well as all occurrences of the C_α conformation were extended for a further 2 ns. C_β and C_δ are equally dominant conformations across the entire ensemble of 60 simulations, in concurrence with the two observed crystal structures C_{1FB7} and C_{1HXB} respectively. However, the drug in the wildtype adopts C_β more frequently whilst in all mutants it prefers C_δ.

and C_δ. The L90M and G48V/L90M mutants favoured C_δ with a substantial concomitant reduction in the frequency of occurrence of C_β and exhibited an increased occurrence of C_α.

The L90M system exhibited a marked increase in the persistence of C_α with only one out of five occurrences evolving into another conformation during the extended 2 ns of simulation. The two dominant conformations of the P2 subsite across all systems were C_β and C_δ, which both had a frequency of adoption in the entire ensemble set of 60 simulations of 0.35 each. The overall frequencies of adoption for C_α, C_γ and C_ε across the entire ensemble were 0.13, 0.07 and 0.10 respectively.

The degree of stability exhibited by each conformation provides some insight into the topology of the energetic landscape in which they are separated. The lack of conformational transitions post-equilibration, in any of the systems, coupled with the existence of stable hydrogen bonds characterising each conformation has several consequences. Firstly it allows the assumption to be made that all trajectories exhibiting conformations C_β - C_ε would have remained in their respective states were they extended for a further 2 ns. As all instances of conformation C_α were explicitly extended, the frequency distribution of conformations shown in Figure 4.9 can then be treated as a consistent distribution. Given that transitions between conformations can occur, the frequency distribution of a set of conformations should match a Boltzmann distribution, allowing us to describe the energy landscape of the conformations; the more frequent a conformation, the more negative the interaction energy between drug and

protease due to it. The lack of transitions observed here indicates that there are significant energetic barriers separating the conformations (excepting C_α) which cannot be overcome within a timescale of 2 ns. This in turn indicates that the frequency distribution observed in our study was not necessarily indicative of the minima of the potential wells characterising each conformation and may instead have been a consequence of the random distribution of initial velocities given to each system.

4.3.3 Drug-Protease Interaction Analysis

In order to assess the validity of the frequency distribution obtained from the ensemble of simulations as a measure of the statistical favourability of each component, we analysed the interaction energy between the protease and saquinavir in the gas phase for each representative conformation for wildtype and all mutant systems. The total gas-phase interaction, $\langle V \rangle$, as well as its decomposition into van der Waals and Coulombic components is shown in Figure 4.10.

The variation of the total interaction energy between protease and saquinavir was much larger across different conformations of the P2 subsite as compared to across different mutant systems (see Figure 4.10(a)). It was therefore not possible to distinguish whether the effect of the mutations studied here was to reduce the binding between the drug and the enzyme. This is not surprising as a thorough differentiation of binding affinity requires inclusion both of the effects of solvation and the changes in configurational entropy. We investigate absolute binding free energy differences further in Chapter 6. Using the method adopted here, it was however possible to clearly differentiate the energetic favourability of each conformation. Across all systems, conformations C_β and C_ε exhibited the most attractive interaction energies varying by over 20 kcal/mol as compared to the other conformations. The order of energetic favourability exhibited by each conformation was conserved across all protease systems (see Figure 4.10(b)). Averaged over the wildtype and mutant systems, this was C_β, C_ε, C_α, C_γ and C_δ which exhibited mean interaction energies of -141.87, -141.12, -118.80, -116.23 and -113.50 kcal/mol respectively. Furthermore the energetic profile for each conformation within each system contrasted significantly from the profile exhibited in the ensemble frequency distribution (see Figure 4.9). Whilst in the wildtype, C_β was both most frequent and most energetically favourable, the occurrence frequency of the other conformations did not follow the energetic profile observed by them. Additionally in all mutant systems, C_δ was the most frequent conformation adopted in the ensemble simulations but was the least energetically favourable. Furthermore, C_ε was energetically very favourable but not frequently adopted in any of the different mutant systems, nor in the wildtype.

Decomposition of the interaction energy into its van der Waals and electrostatic components allowed the source of the variation exhibited between the conformations to be investigated. The van der Waals energies across all conformations in all protease systems were very similar (see Figure 4.10(c)). Conversely, a large variation in the electrostatic component of interaction energy was exhibited across different conformations (see Figure 4.10(d)). The electrostatic profile followed the changes in total in-

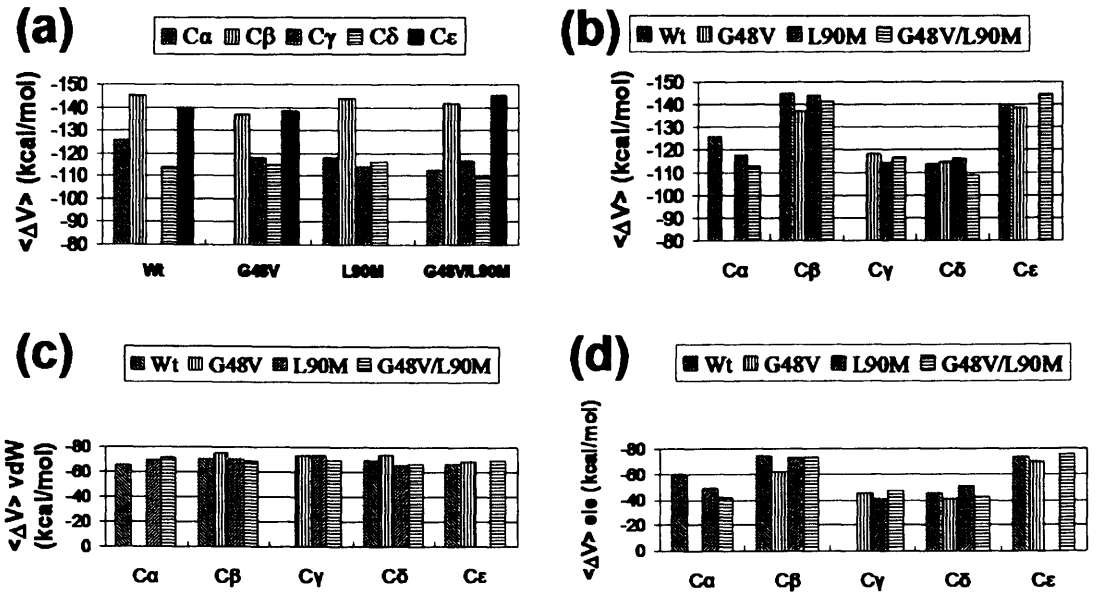


Figure 4.10: Total gas-phase interaction energy, $\langle V \rangle$, between protease and saquinavir decomposed by (a) protease mutant and (b) P2 subsite conformation. (c) The van der Waals component of the interaction energy decomposed by P2 subsite conformation and (d) the electrostatic component of the interaction energy.

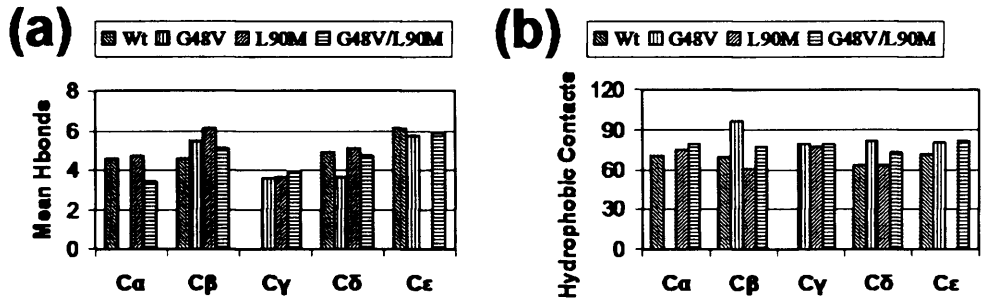


Figure 4.11: (a) Mean number of all drug-protease hydrogen bonds for each protease system across each conformation. (b) Mean number of drug-protease hydrophobic contacts for each protease across each conformation.

teraction energy very closely. Again C_β and C_ε showed significantly increased electrostatic interaction compared to the other conformations. As expected, C_α and C_γ, which had no significant P2 subsite interactions with the protease, showed less relative electrostatic interaction. Interestingly, even though C_δ exhibited relatively strong P2 subsite hydrogen bonds, relatively small electrostatic affinity was observed for it.

Although the electrostatic profile exhibited here was supported by the existence of strong P2 subsite hydrogen bonds and especially of several hydrogen bonds in the case of C_ε, the P2 subsite H-bond network alone was not enough to explain the large increases in electrostatic interaction observed for the C_β conformation as well as the significant comparative reduction observed for C_δ. This indicated that the various conformations caused additional indirect changes in the H-bond network between the rest of the drug and the protease.

To investigate this further, the mean number of all drug-protease hydrogen bond interactions was analysed (see Figure 4.11(a)). C_β and C_ε again exhibited a greater number of mean drug-protease H-bonds (between 1 and 2) than any other conformation. Furthermore, even though C_β and C_δ are characterised by only one hydrogen bond associated with the N^4 atom of the P2 subsite, the additional mean number of hydrogen bonds for the entire drug supported the increased electrostatic interaction energy difference observed between these conformations. Similarly the mean number of drug-protease hydrophobic contacts was also analysed (see Figure 4.11(b)). This profile followed the van der Waals contribution to the electrostatic interaction energy closely; there was insignificant difference across all conformations and little variation was observed between different protease systems for all conformations except C_β. The variation in C_β was due to an outlying result for the G48V system in which there was a significant increase in hydrophobic contacts.

4.3.4 Decomposition of Hydrophilic and Hydrophobic Interactions

Active Site Sub-Region Decomposition

We investigated the decomposition of hydrophilic and hydrophobic interactions between the subsites of the drug and selected sub-regions of the protease for the various P2 subsite conformation-exhibiting simulations across all protease systems. The active site of the protease was decomposed into separate spatial sub-regions (see Figures 4.12(a) and 4.12(b)) in a way that encompassed all possible hydrogen bonds and hydrophobic interactions between the drug and the active site.

The 'Flap' sub-region consists of residues I47 to I50 and I147 to I150 that make up the inner strands of the flaps as well as the tetrahedrally co-ordinated water molecule between the flaps and the inhibitor, the 'Asp' sub-region is composed of the catalytic aspartic acid dyad. The 'Outer' sub-region consists of residues G27 to G30, R108, L123, G127 to G130, R8 and L23. Finally the 'Wall' sub-region of the active site consists of V32, P81, V82, I84, V132, P181, V182 and I184. The 'Flap' and 'Outer' sub-regions therefore contain a mixture of hydrophilic and hydrophobic residues, whilst the 'Asp' sub-region is hydrophilic and the 'Wall' sub-region entirely hydrophobic. The drug subsites were also partitioned as shown in Figure 4.1(b).

Figure 4.13(a) shows the decomposition of the mean number of H-bonds (μ) between the inhibitor and each of the active site sub-regions across all conformations and across all protease mutants. The P2 subsite hydrogen bond contribution associated with the N^4 atom to each of these sub-regions is also shown (black). The Wall sub-region (green) did not exhibit any hydrogen bonds with the inhibitor.

For conformations C_α, C_γ and C_δ, the Flap (red) was the dominant sub-region which mediated H-bond interactions with the inhibitor. Furthermore, these conformations exhibited similar and relatively small Outer (orange) and Asp (blue) hydrogen bonds for each of which $\mu \sim 1$ across all protease systems except L90M, for which there was a marginal increase in Outer H-bonds. Conformation C_δ exhibited the largest Flap contribution in all protease systems again except L90M for which the Flap contribution for C_β was anomalously large. Furthermore, the increased H-bond contribution observed in C_δ compared to other conformations ($\Delta\mu \sim 1$) was entirely due to the additional H-bond formed between the N^4 atom of the P2 subsite and the flaps. The large increase in mean H-bonds observed in conformation C_ε (see Figure 4.11(a)) was entirely due to the direct increase in hydrogen bonding conferred by the P2 subsite bonding with the hydrophilic well contained in the Outer sub-region and not due to further alteration of the hydrogen bond network between the inhibitor and the Asp and Flap sub-regions. Conformation C_β was the only conformation across all systems to not only directly confer an additional hydrogen bond with the Asp sub-region through direct P2 subsite interaction, but also to indirectly alter the hydrogen bond network to allow additional non-P2 subsite hydrogen bonds to form with the catalytic dyad. Visual inspection confirmed that the indirect increase in Asp H-bonding was due to the reorientation of the N^3 atom of the P1 subsite to form a hydrogen bond with the OD2 atom of residue

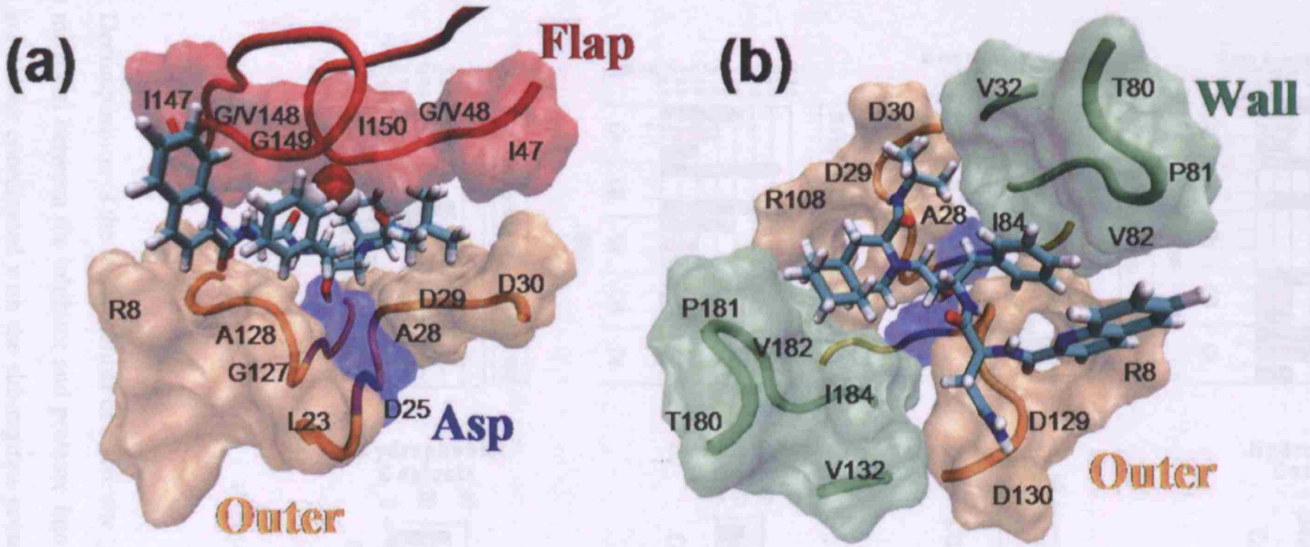


Figure 4.12: Schematic diagrams of active site decomposition into several distinct sub-regions from (a) side-view and (b) top-down view: Flap (residues 47 to 50 and 147 to 150 as well as the tetrahedrally coordinated water molecule between the flaps and the inhibitor), Asp (D25 and D125 residues), Wall (residues V32, P81 to I84, V132 and P181 to I184) and Outer (residues R8, L23, G127 to D130, R108, L123 and G27 to G30). For clarity, the Wall sub-region is not shown in (a), nor is the Flap sub-region shown in (b).

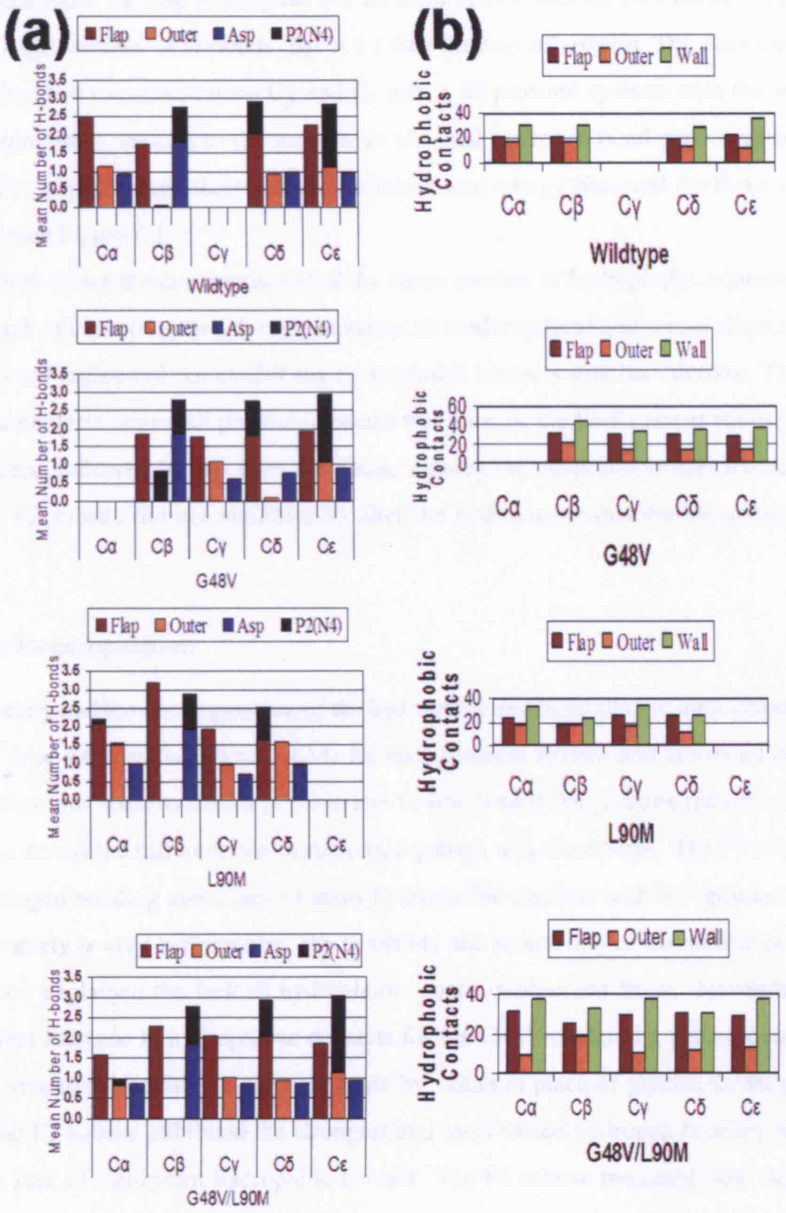


Figure 4.13: Decomposition of the mean number of active site (a) hydrogen bonds (μ) and (b) hydrophobic contacts mediated between the inhibitor and protease into Flap, Outer, Asp and Wall sub-regions. Interactions are colour coordinated with the sub-regions presented in Figure 4.12. Hydrophilic interactions mediated by the tetrahedrally coordinated water molecule between the flaps and the inhibitor are included in the 'Flap' sub-region component. The sub-region contribution of the hydrogen bonds associated with the N^4 atom of the P2 subsite are also included (black).

D125. Furthermore, only for conformation C_β did the mean number of hydrogen bonds with the Asp sub-region exceed those for Flap sub-region. For all other conformations, the Flap sub-region exhibited a significantly larger number of H-bonds ($\Delta\mu > 1$) than the Asp sub-region. The consistent increase of two hydrogen bonds for conformations C_β and C_ε across all protease systems with the Asp and Outer sub-regions respectively, leading to the increase in the total hydrogen bond pattern with the inhibitor (see Figure 4.11), explained the large increase in interaction energy observed for these conformations over the others (see Figure 4.10).

Figure 4.13(b) shows the decomposition of the mean number of hydrophobic contacts between the inhibitor and each of the active site sub-regions across all conformations and across all protease mutants. The Asp (blue) sub-region did not exhibit any hydrophobic contacts with the inhibitor. The distribution of hydrophobic contacts across all protease systems was similar, the Wall (green) sub-region exhibited the most contacts, followed by the Flap and Outer sub-regions respectively and the various conformations of the P2 subsite did not significantly alter the hydrophobic distribution across the different sub-regions.

Drug Subsite Decomposition

Finally, we investigated the decomposition of the hydrophilic and hydrophobic interactions with respect to each of the drug subsites (see Figure 4.14) for each protease system and across all conformations. Even though there was some variation in the extent to which each drug subsite interacted with the protease active site across the different conformations, a pattern was discernible. The P3 subsite exhibited both weak hydrogen bonding and a lack of many hydrophobic contacts with the protease. Even though the quinoline moiety is very hydrophobic, the flexibility and rotatability of the subsite as confirmed by visual inspection explained the lack of hydrophobic contacts observed for it. Interestingly however, there is a distinct increase in hydrophobic contacts for the G48V-containing mutants compared to the other systems, due to the additional contacts made by valine in place of glycine in the protease flaps. As expected the P2 subsite exhibited the strongest and most varied hydrogen bonding with the active site and also a lack of significant hydrophilic contact. The P1 subsite mediated both strong hydrogen bonding and hydrophobic interaction with the active site. The hydrogen bonding was due to the hydroxyethylene moiety bonding with the catalytic dyad, whilst the hydrophobic contacts were conferred by the phenyl group interacting with both the Flap and the hydrophobic Wall sub-regions. The P1′ and P2′ subsites, as expected exhibited strong hydrophobic interactions. However, whilst the oxygen atom O^1 of the P1′ subsite was coordinated to the inter-flap-inhibitor water molecule, thus exhibiting one hydrogen bond, no hydrogen bonding was exhibited by the P2′ subsite in any of the protease systems.

Interestingly, even though the central subsite, P1, exhibited both strong hydrophilic and hydrophobic interactions, it was flanked on either side by subsites that interacted predominantly in either a hydrophilic or hydrophobic way but not in a significant combination of both. Furthermore, whilst the P2′ end subsite

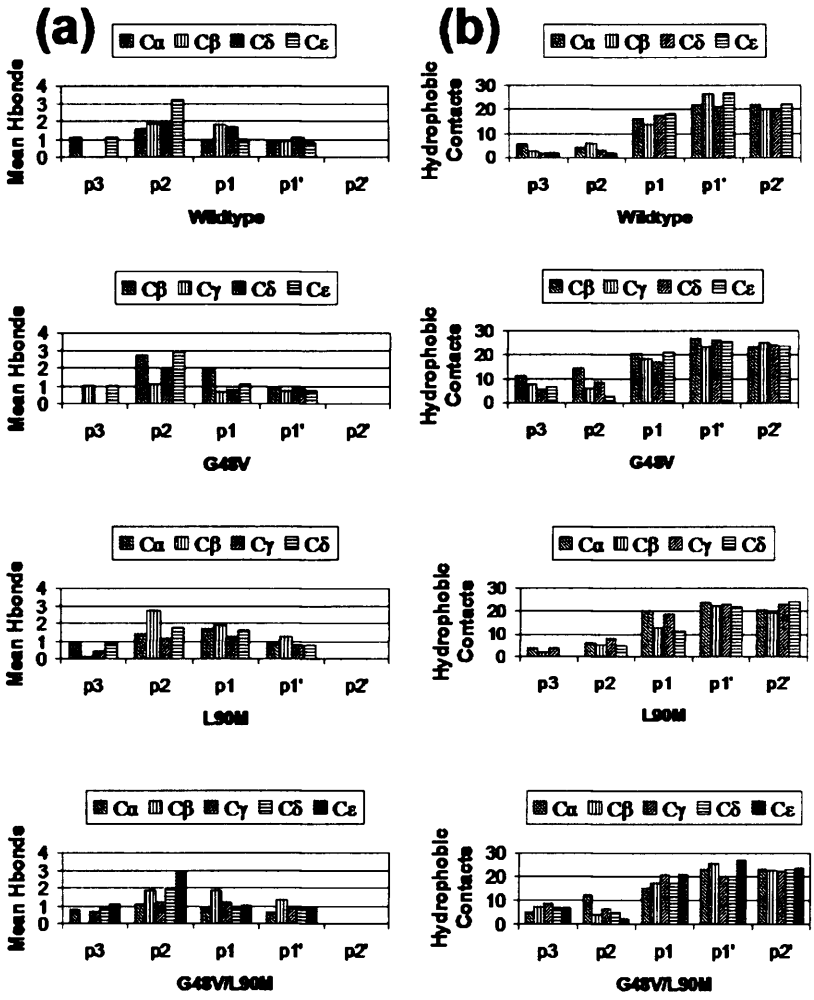


Figure 4.14: Decomposition of the mean number of active site (a) hydrogen bonds and (b) hydrophobic contacts mediated between the inhibitor and protease across the drug subsites P3-P2′.

interacted strongly with the active site, the interaction was not matched by that of the P3 subsite at the other end of the drug.

4.4 Discussion

In order to investigate the multiplicity of the different subsite conformations of an inhibitor that may exist in the active site of HIV-1 protease as well as the diverse role of their resulting interactions and to provide insight into how mutations may effect these conformations, we performed multiple simulations of the protease bound to the inhibitor saquinavir for each of the wildtype, the G48V and L90M single mutants and the G48V/L90M double mutant systems in explicit water.

We focussed on the multiplicity of the conformations of the P2 subsite of saquinavir, which is identical to an asparagine side-chain and plays a crucial role in substrate specificity, being found in four of the ten cleaved substrates. We are able to draw several conclusions from our study. Firstly, alongside the two conformations that show similarity to the two existing crystal structures, the conformational landscape of the P2 subsite of saquinavir is characterised by three additional conformational minima, stable over a 2 ns timescale and which are distinct in terms of dihedral angle rotations along the P2 subsite as well as mutually exclusive hydrogen bonds with different regions of the active site (see Figure 4.7) and which induce significant variations in the drug-protease interaction energy (see Figure 4.10). Although not explicitly calculated here, the lack of conformational transitions observed in our simulations imply that these conformational minima are separated by significant energetic barriers, not readily surmountable over a 2 ns timescale. Furthermore, the similarity of the crystal structure conformations with two of these minima implies that the additional conformations shown here are physically possible and may be observed experimentally if more crystal structures of the complex existed.

Secondly, the significant differences in the drug-protease hydrogen bond network exhibited by the various conformations (see Figures 4.11(a) and 4.13)(b), which explain the significant variation in the observed drug-protease interaction energies (see Figure 4.10), coupled with the observation that there are no significant conformational transitions post-equilibration, over a 2 ns timescale, have potential implications for the subsequent dynamics of the inhibitor. Lack of conformational transitions implies that the adoption of a particular conformation may persist well beyond the 2 ns timescale investigated here and subsequently allow differential interactions to be exhibited between the drug and the protease. Alongside the absolute strength of the drug-protease interaction, the spatially decomposed network of interactions with various sub-regions of the active site may then be a key factor determining the differential behaviour over a longer timescale. The drug may not be able to climb out of the energetic well of one conformation into another before other drug-protease interactions are able to substantially alter the phase-space available to the system.

For example, only in one conformation (C_β) is significantly greater hydrogen bonding observed with

the catalytic dyad than with the flaps. In all other conformations, the bonding with the dyad is reduced by two hydrogen bonds and is significantly smaller than the bonding with the flaps. As the flaps have previously been shown to be very flexible in the apo-protease [132] and exhibit larger flexibility than the active site even in the presence of saquinavir [242], whilst the dyad has been shown to be very inflexible, the large anisotropy in the ligand-protease hydrogen bonding between the flap and the dyad in the other conformations and especially the significant reduction in the Asp hydrogen bonding may adversely influence the drug-protease interaction. Since the Asp and Flap sub-regions diametrically oppose each other with respect to the inhibitor (see Figure 4.12), increased coupling to the flaps is likely to induce motion of the inhibitor away from the active site centre, especially given the well characterised motion of the flap towards open conformations [142, 252]. This supports recent studies which have suggested that interactions with the catalytic residues should be given preference over total binding affinity when designing drugs [253].

Furthermore, as the hydrophobic contact profile is left largely unaltered (see Figures 4.11(b) and 4.13(b)) across different conformations, such differentially induced motion would be largely initiated by the conformational differences in the P2 subsite and the direct and indirect changes in the hydrogen bond network induced by them. In Chapter 5, we investigate the dynamics of saquinavir in the active site over a longer timescale for the same protease systems studied here and observe that different P2 subsites do indeed lead to significantly different dynamical behaviour of the inhibitor over a timescale of 25 ns. The effects of the G48V and L90M mutations on flap dynamics in the presence of saquinavir are also studied therein.

The two dominant and equally frequent conformations C_β and C_δ adopted in our overall ensemble of simulations concur with the crystal structures C_{1HXB} and C_{1FB7} respectively, although C_β varies with C_{1FB7} due to additional hydrogen bonds with the catalytic dyad. However, in the wildtype the P2 subsite adopts conformation C_β more frequently in our ensemble of simulations, whilst C_δ is more frequent in all mutant systems. Certainly, C_β is more consistent with the orientation of the P2 subsite in natural substrates [129] and implies that in saquinavir bound to the wildtype protease it accesses the intended S2 pocket enclosed by residues 125 to 128 of the active site [241].

The differences in the frequency distribution across different mutant systems (see Figure 4.9) cannot be interpreted as a distribution of the most thermodynamically favoured states. If multiple conformational transitions had been observed in each run, then a statistical distribution of the overall conformational frequency would have been an appropriate measure of conformational favourability. From an energetic perspective (see Figure 4.10), only in the wildtype does the most frequent conformation (C_β) match the most energetically favourable. In all mutant proteases, even though conformation C_δ is most frequently adopted, it is also the least energetically favourable. The differences between the ensemble frequency distribution and the drug-protease interaction energy distribution therefore imply that the frequency distribution observed here may be dependent on the proximity of the initial point in phase space,

given to each trajectory by the randomly assigned initial velocity configuration, to the conformational minimum which it adopts.

The large frequency of adoption of C_δ in G48V containing mutants may be explained by the additional conformational restraints placed on residues 48 and 148. The substituted valine in the G48V mutant causes increased steric hindrance, preventing complete rotation of the side-chain into the flaps and constraining the quinoline moiety of saquinavir at subsite P3 through increased hydrophobic interactions (see Figure 4.14(b)). This in turn limits the conformational freedom of the backbone carbonyl oxygen of residue 148 involved in hydrogen bonding to the P2 subsite. As the starting conformation C_{1HXB} is already close to the C_δ minimum, such additional constraints would lead to an increase in the energetic barrier, reducing the probability of transition into another conformation. The increased frequency of adoption of C_β in the wildtype may be explained by the lack of such constraints which would lead to a reduced transitional barrier. Furthermore, the strong hydrophilic attraction of the amide end of the P2 subsite to the dianionic catalytic dyad would explain why upon adoption of C_α, further rotation occurs for most systems into C_β. Therefore, even though the existence of multiple conformations and their differential effects are well described in this study, we are unable to fully differentiate the effects of mutations in terms of an alteration in the statistical frequency of adoption of each conformation over this timescale.

There is significant debate over the protonation state of the dyad even though several studies have been undertaken regarding it [147, 156]. Whilst at physiological pH, the catalytic dyad should be dianionic, optimal catalytic efficiency occurs at a slightly acidic pH when the protease may or may not be in the monoprotonated state. Altering the protonation state may therefore alter the conformational space accessible to the P2 subsite.

The multiplicity of the conformations exhibited by the P2 subsite as well as the profile of the interactions of each drug subsite with the protease active site (see Figure 4.14) has significant implications for drug design. Firstly, the P3 subsite only weakly interacts with the active site compared to other non-central subsites such as P1′ and P2′. Whilst it may be advantageous to preserve these other subsites, modifications of the P3 subsite may increase interactions and enhance binding. Interestingly the same P1′ and P2′ subsites have been used in the design of nelfinavir. Second generation inhibitors containing subsites with additional rotatable bonds have been shown to adapt better to mutant proteases [211]. Controlled subsite rotatability is therefore an attractive prospect for the design of better inhibitors. Understanding the extent of rotational flexibility of a subsite is essential however, in limiting the adverse effects which may arise from unforeseen conformations being adopted. If the conformational landscape available to an inhibitor in the active site of the protease is not sufficiently mapped out, then mutations in the protease may be able to take advantage of certain conformations to confer additional resistance through the alteration of drug-protease interactions. Conversely, understanding the effects of certain drug-resistant mutations on the accessibility of the subsite conformations of prospective in-

hibitors would allow for the rotational flexibility of such subsites to be optimised to cater both for the wildtype and a set of known mutants.

CHAPTER 5

Insights into a Mutation-Assisted Lateral Drug Escape Mechanism from the HIV-1 Protease Active Site

MOLECULAR mechanisms regarding the association of ligands with proteins are difficult to determine experimentally. In Chapter 1 we discussed experimental methods that are used to probe both the thermodynamics and kinetics of protein-ligand association. However, whilst such thermodynamic and kinetic properties are determinable, experiments that yield such information do not confer understanding about the molecular mechanisms along the pathway of a binding event. Molecular binding and dissociation requires some degree of conformational and interactional adjustment between the binding species. The energy required to do this is a barrier to the binding and contributes to the 'activation energy' of the association and is thus ultimately related to the kinetic rate of the reaction. Unfortunately, insights into kinetic differences arising between a set of varying protein-ligand binding reactions are limited to providing differences in the rate constants of each of the reactions, but are not able to explain why such differences occur at the molecular level.

In theory, molecular dynamics simulations are able to 'fill the gap' by providing such insight. Conventionally, the use of molecular dynamics to tackle such a problem has been intractable, due to the fact that many binding events happen over a timescale unapproachable by fully atomistic simulations (between the μs to ms). Conversely, if the atomistic description of a molecular system is replaced by a more coarse-grained representation, the exact cause of the kinetic barrier to the progress of a reaction becomes more difficult to pin down. Fortunately, the ongoing increase of computational power coupled with the advances in high performance computing, discussed in Chapter 2, allow longer atomistic simulations to be achieved.

In the context of the HIV-1 protease, the effect of mutations on the binding affinity of inhibitors has been discussed in Chapter 3 and is studied further in Chapter 6. However, a thermodynamic description of resistance should be complemented by a molecular understanding of the changes in the

kinetic properties of inhibitor dissociation. In this chapter, we study the differential interactions of an inhibitor bound to the wild type and several mutant forms of HIV-1 protease. The timescale for complete dissociation of an inhibitor from the HIV-1 protease is still intractable by conventional molecular simulation. Our aim however, is to explore the deviations of an inhibitor from the bound state through the use of temporally extended fully atomistic molecular simulations and as such to provide insight into the kinetic mechanisms of drug resistance at the molecular level. The chapter is organised subsequently into a 'Background' section, in which we outline the specific context of the study, a 'Methods' section, highlighting the specific computational protocols that were implemented, a 'Results' section in which we report our findings and finally a 'Discussion' section.

5.1 Background

Due to its key role in the cleavage and subsequent maturation of the structural (Gag) and enzymatic (Pol) proteins from their precursors, the protease has been a key target for structure-based anti-retroviral inhibitors [166, 241]. Unfortunately, the high replication rate of the virus and low fidelity of the reverse transcription process have led to the proliferation of several resistant strains that follow emerging mutational patterns [170, 174, 176, 177].

The loss in binding affinity of inhibitors binding to drug resistant mutants of HIV-1 protease as compared to the wildtype has been very well studied experimentally [15, 202, 204] and computationally [218, 245, 254]. However, such calculations do not provide information about the kinetic mechanism of drug association or dissociation, nor can they alone explain the variation in the kinetic role of mutations in causing resistance.

For example, whilst it is likely that some active site mutations like V82A and I84V cause resistance through direct steric hindrance, other mutations not in the active site cause resistance through an alteration in the dynamical properties of the enzyme [139, 217]. Furthermore, for some such mutations, although structural differences between proteases remain small, significantly different clinical behaviour is observed [173].

The role of the flaps in the catalytic mechanism of HIV-1 protease has been extensively studied. Ligands bind to the protease in a two-step binding process, firstly forming a loose complex, with the flaps in an open conformation and secondly with the flaps securing the ligand in a closed conformation [255]. Previous studies of the many crystal structures reported for HIV protease complexes, have revealed the characteristic flexibility of the flaps together with the stability of the catalytic region [132]. Furthermore, the crystal structures of apo-proteases are almost entirely in a third, semi-open flap conformation, whilst ligand-bound structures are predominantly in a closed conformation.

Protease flexibility, stability and flap motion have also been extensively studied using computational techniques such as molecular dynamics [136–138]. Recent studies on both the free and inhibitor-bound

enzyme suggest the predominance of the semi-open flap conformation in the unliganded protease which can repeatedly open and close, whilst showing that the flaps remain stably closed when an inhibitor is bound [142]. This is in good agreement with previous NMR studies [143, 144] that have suggested equilibria between open, semi-open and closed forms of the flaps. Interestingly, a recent crystal structure of an unbound multi-drug resistant (MDR) protease with flaps in an open conformation was found to revert back to a semi-open form in molecular dynamics simulations [135] and it was suggested that such an open form was due to crystal packing effects. Molecular dynamics simulations have also provided insights into the mechanism of the reversal of flap handedness upon binding of a substrate, which varies between closed and semi-open conformations [252, 256].

The effect of drug resistant mutations on the dynamics of the protease has also been studied. Previous, fully atomistic simulations have shown the increase in flexibility of the unliganded form of the drug resistant V82F/I84V double mutant over the wildtype [114], with the flaps occupying a semi-open conformation more often in the mutant. Coarse-grained Brownian dynamics simulations have also shown that the effect of some mutations is to reduce the frequency of flap opening with respect to the wildtype and thus contributing to a decrease in the association rates of inhibitors [141].

However, whilst significant work has now been done in determining the flap-associated mechanisms of drug-binding to the protease in an open flap conformation, the reverse process of mechanistic drug dissociation remains poorly understood, especially at the atomistic level. Due to the stabilisation of the flaps in a closed conformation upon inhibitor binding, the timescale to observe any significant drug deviation away from a bound state coupled to an opening of the flaps has remained beyond the scope of molecular simulation [142].

Here, we perform temporally extended, fully atomistic molecular simulations, each of 25 ns with explicit solvent, of the wildtype protease and three drug resistant mutants (G48V, L90M and G48V/L90M) bound to the inhibitor saquinavir. We provide insights into the first stages of a lateral drug dissociation mechanism from the active site of the protease, following reversion of the flaps into a semi-open conformation. Furthermore, we explore the differential interactions in each protease mutant compared to the wildtype and thus provide insights into the mechanistic basis of drug resistance conferred by the G48V mutation due to enhanced inhibitor coupling with the highly flexible flaps of the protease. Our analysis includes the evolution of the conformations of the P2 subsite of saquinavir, discussed in Chapter 4 and its initial role in precipitating differential active site interactions.

The postulated expulsion mechanism, which we investigate through the use of steered molecular dynamics simulations [55], is likely not to require the full opening of the flaps and the subsequent alteration of previously proposed dissociation rates for inhibitors from wildtype and mutant proteases [15] is therefore discussed.

5.2 Methods

5.2.1 Initial Preparation, Equilibration and Production Runs

The initial preparation of all four systems as well as the minimisation and equilibration protocols employed are fully described in Chapter 4.

Each of the four simulations performed in this study were extensions of singular runs in the isothermal-isobaric ensemble (NPT) selected from each simulation ensemble in the study performed in Chapter 4. These were R2, R0, R0 and R0 for the wildtype, G48V, L90M and G48V/L90M systems respectively. As mentioned previously, the first nanosecond of simulation in the NPT ensemble, in which the volume of the water box adjusted to re-establish and maintain a uniformly dense water box, was taken as the final phase of equilibration. This resulted in a total of 2.23 ns of equilibration for the wildtype and G48V systems and 2.28 ns of equilibration for the L90M and G48V/L90M systems.

The output coordinates from each of the four protease systems were then used as the starting points for all subsequent production runs, each of which continued in the NPT ensemble for a further 24 ns. Coordinate trajectories were recorded every 1 ps throughout all equilibration and production runs. In total, over 100 ns of simulation was achieved and performed under conditions of optimal computational efficiency using NAMD2 [34], with a wall-clock rate of approximately 8 hours/ns, using 30 processors on a 512 processor SGI Altix at CSAR, University of Manchester, UK, 32 processors (1 node) at the UK national HPCx facility, Daresbury, and 32 processors on the TeraGrid cluster at NCSA. These simulations also made use of 32 processors of the Leeds and Oxford compute nodes of the UK National Grid Service.

5.2.2 Post-Production Analysis and Steered Molecular Dynamics

The simulations were analysed using a range of methods. Root mean squared deviation (RMSD), distance vector and angular displacement analyses were calculated using 'tcl' scripts in VMD. The criteria used for an instantaneous hydrogen bond was a donor-acceptor distance ≤ 3.5 Å and a donor-hydrogen-acceptor angle $\geq 150°$. Instantaneous hydrogen bonds were calculated each 1 ps and averaged over a time-window of 100 ps (see Figure 5.9). Hydrophobic contacts were also calculated using the same method as that for the hydrogen bond analysis. The criteria for the existence of instantaneous hydrophobic contacts was an atom-atom distance ≤ 3.5 Å.

Radii of gyration and radial distribution functions were calculated using the PTRAJ module in the AMBER 9 software package [53]. PTRAJ was also used for implementation of principal component analysis (PCA). For PCA, the backbone protease atoms (C_α, C and N) as well as the non-hydrogen drug atoms were used for the analysis. Production trajectories of all systems were combined and fitted to the 1HXB crystal structure as an initial reference structure; the covariance matrix and principal eigenvectors

were calculated using this combined trajectory, making direct comparison between systems possible over a consistent eigenvector set. Principal projections for each system were superimposed every 250 snapshots and represented schematically (see Figure 5.8) using the IED package [54] interfaced with VMD.

Steered molecular dynamics (SMD) simulations were implemented for each protease system using NAMD2, and made use of 32 processors on the Leeds and Oxford compute nodes of the UK NGS, from the last coordinates of the unsteered MD simulation. The steered 'dummy' atom was attached via a force constant of k = 10 kcal/mol/Å^2 to all non-hydrogen drug atoms and pulled with a steered velocity of v = 0.01 Å/ps (see § 2.5.1). These values of force constant and steered velocity corresponded to a stiff spring in the drift regime [56]. Furthermore, the velocity was slow enough to allow relaxation of the solvent in response to steering. To prevent translation and rotation of the protease molecule upon application of the steering force, several C_α atoms were held fixed, specifically those of the N- and C-termini residues 1, 99, 100 and 198 as well as residues 30, 108, 123 and 181 at the back-end of the protease. The direction of steering was determined by the vector with origin at the C_α atom of residue R108 pointing in the direction of the C_α atom of residue R8; the 'dummy' atom was pulled to a distance of 15 Å from its starting position, corresponding to a steered simulation time of 1.5 ns. Centre of mass displacement and SMD force values were output every 10 timesteps and all other MD parameters were kept the same as those used in the NPT ensemble, described in Chapter 4.

5.3 Results

5.3.1 Structural Flexibility

As a preliminary indication of global backbone flexibility, we measured both the backbone root mean squared deviation (RMSD) and the radius of gyration of each of the four systems from their original starting structures across the entire trajectory (see Figure 5.1).

Over the first 5 ns of simulation, no significant difference in RMSD is discernible between any of the systems with values ranging from 1.25 Å to 1.5 Å. However further RMSD analysis shows that the L90M mutant is more flexible than the other three systems manifesting the largest deviation from the original structure over the rest of the simulation. There is also little change in the radius of gyration of each of the four systems over such a timescale, with mean values of 17.56 ± 0.37 Å, 17.58 ± 0.37 Å, 17.65 ± 0.39 Å and 17.53 ± 0.39 Å for the wildtype, G48V, L90M and G48V/L90M systems respectively. The results also show that the L90M mutant exhibits the largest flexibility in accord with the RMSD analysis. The backbone root mean squared fluctuation (RMSF) relative to the average structure across the whole of the production run for the wildtype, G48V, L90M and G48V/L90M systems is 0.93 ± 0.14 Å, 0.84 ± 0.10 Å, 1.01 ± 0.20 Å and 0.87 ± 0.12 Å respectively, also supporting the slightly increased

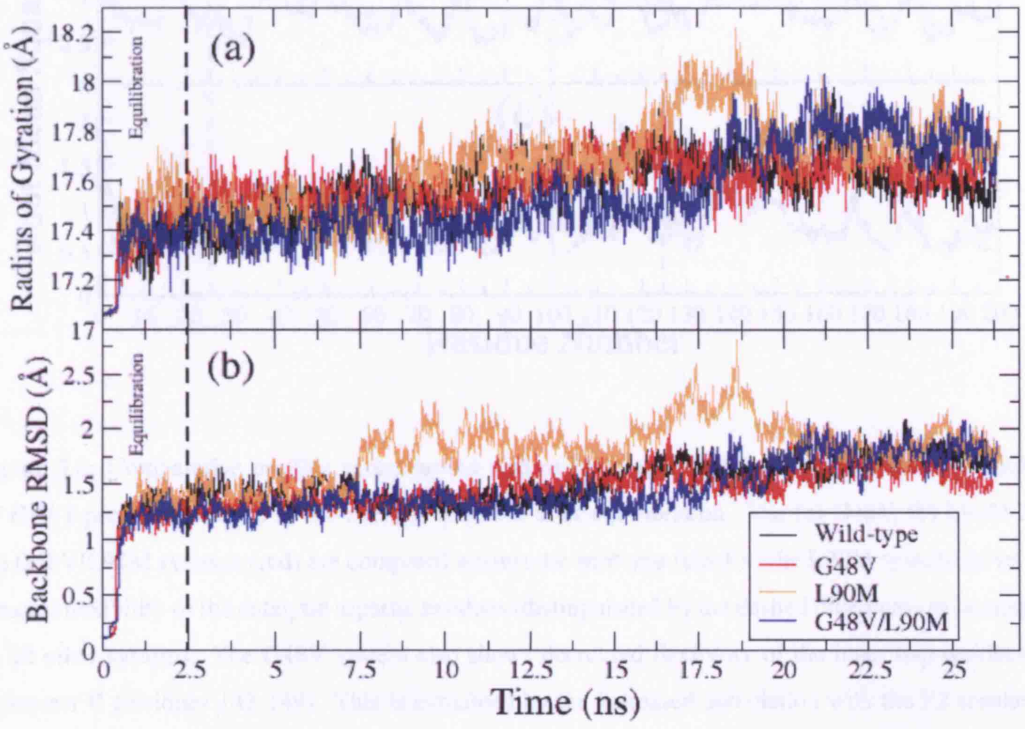
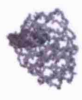


Figure 5.1: (a) The radius of gyration and (b) RMSD of backbone atoms of HIV-1 protease relative to the crystal structure, for the wildtype (black), G48V (red), L90M (orange) and G48V/L90M (blue) systems. The same colour scheme is used throughout this paper when all four systems are compared. The L90M mutant shows slightly larger flexibility than the other systems across the timescale of 25 ns.



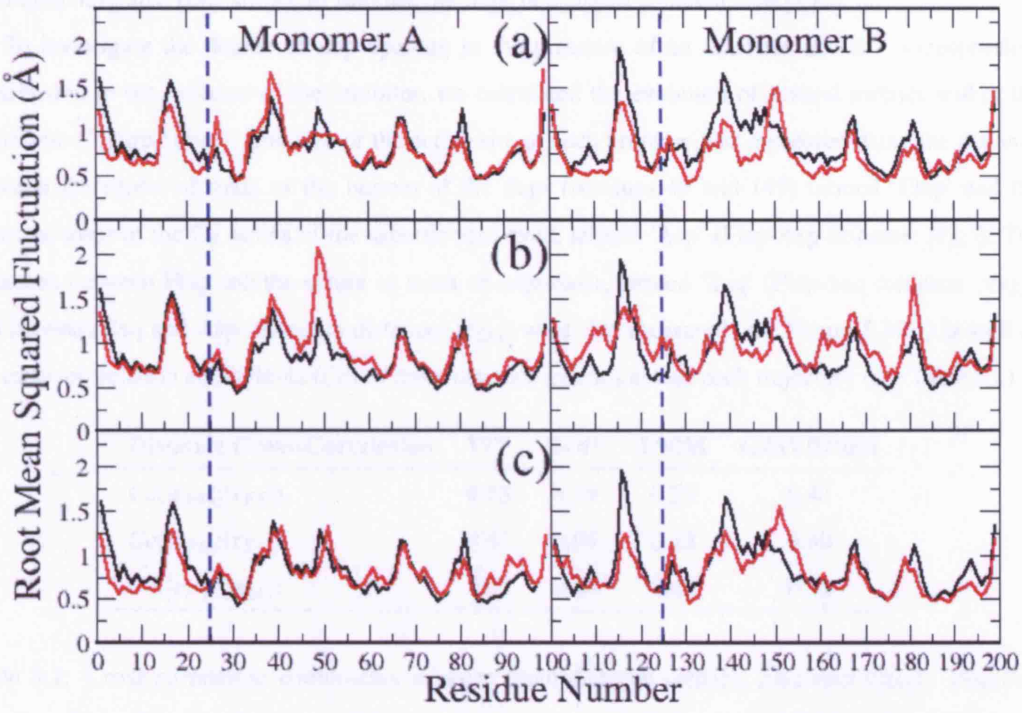


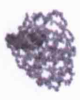
Figure 5.2: Comparative profiles versus amino acid residue number of the RMSF of backbone atoms of HIV-1 protease, relative to the average structure after equilibration. The (a) G48V, (b) L90M and (c) G48V/L90M systems (red) are compared against the wildtype (black). The L90M system shows increased flexibility of the catalytic aspartic residues (distinguished by the dashed blue lines) as compared to all other systems. The G48V system also shows decreased flexibility of the inner flap residues of monomer B (residues 143-149). This is explained by the increased association with the P2 subsite of saquinavir, which exhibits conformation C_δ here.

flexibility of the L90M mutant.

In addition to changes in the global flexibility of the enzyme, the specific change in flexibility across each of the residues was also calculated. Figure 5.2 shows the RMSF relative to the same global average structure for the backbone atoms of each residue. All systems showed similar fluctuations in their catalytic residues except for that of monomer B in the L90M mutant system, which showed an RMSF of 1.08 Å as compared to 0.64 Å in the wildtype.

5.3.2 Coupled Flap and Inhibitor Dynamics

Previous studies on HIV apo-proteases have shown that some mutations alter the equilibrium between more open and more closed conformations of the flaps, facilitated by curling of the flap tips [114, 257].



Inhibitors have also been shown to stabilise the flaps in a closed conformation [140].

To investigate the degree of flap opening in the presence of an inhibitor and the corresponding relationship to the position of the inhibitor, we calculated the evolution of several metrics within the active site (Figure 5.3(a)). The size of the active site of each protease was measured using the distance between the centre of mass of the bottom of the flaps (residues 49 and 149) termed 'Flap' and the centre of mass of the C_β atoms of the aspartic acid dyad, termed 'Asp' (Flap-Asp distance: $|\mathbf{r}_{FA}|$). The distances between Flap and the centre of mass of saquinavir, termed 'Saq' (Flap-Saq distance: $|\mathbf{r}_{FS}|$), and between Saq and Asp (Saq-Asp distance: $|\mathbf{r}_{SA}|$) were also measured (see Figure 5.3(b)) as well as the cross-correlation coefficients (Cc) of these distance metrics across each trajectory (see Table 5.1).

Distance Cross-Correlation	**WT**	**G48V**	**L90M**	**G48V/L90M**
$Cc(\|\mathbf{r}_{FA}\|:\|\mathbf{r}_{FS}\|)$	**0.85**	0.39	0.25	0.41
$Cc(\|\mathbf{r}_{FA}\|:\|\mathbf{r}_{SA}\|)$	0.41	**0.91**	0.13	**0.90**
$Cc(\|\mathbf{r}_{FS}\|:\|\mathbf{r}_{SA}\|)$	0.01	0.44	0.45	0.22

Table 5.1: Cross-correlation coefficients between three distance metrics, Flap-Asp ($|\mathbf{r}_{FA}|$), Flap-Saq ($|\mathbf{r}_{FS}|$) and Saq-Asp ($|\mathbf{r}_{SA}|$), across all four protease systems. A high value of $Cc(|\mathbf{r}_{FA}|:|\mathbf{r}_{FS}|)$ indicates tight drug-coupling to the aspartic acid dyad, whilst a high value of $Cc(|\mathbf{r}_{FA}|:|\mathbf{r}_{SA}|)$ indicates tight drug-coupling to the flaps. Only in the wildtype is the drug significantly coupled to the aspartic acid dyad. The G48V-containing systems both exhibit tight coupling of the drug to the flaps. $Cc(|\mathbf{r}_{FS}|:|\mathbf{r}_{SA}|)$ is not significant in any system as the point 'Saq' does not necessarily lie on the Flap-Asp vector.

In our simulations, the drug manifests stable binding to the wildtype protease at a distance of 6.5 Å from the aspartic acid dyad, whilst also stabilising the flaps at 11 Å. A low cross-correlation coefficient between the Flap-Asp and Saq-Asp distances, $Cc(|\mathbf{r}_{FA}|:|\mathbf{r}_{SA}|) = 0.41$, indicates that the flap motion is decoupled from the motion of the drug whilst a high cross-correlation coefficient between the Flap-Asp and Flap-Saq distances, $Cc(|\mathbf{r}_{FA}|:|\mathbf{r}_{FS}|) = 0.85$, confirms that the drug is coupled tightly to the aspartic acid dyad. Conversely, the G48V mutant shows tight coupling between the motion of the drug and the flaps ($Cc(|\mathbf{r}_{FA}|:|\mathbf{r}_{SA}|) = 0.91$), demonstrated clearly after 12 ns of simulation, where a discrete change in the Saq-Asp distance (orange line) precedes that of the Flap-Asp distance (black line) by approximately 50 ps. In the subsequent 13 ns of the simulation the drug moves nearly 3 Å away from the active site centre. The L90M mutant shows the most fluctuation in flap dynamics, although no stable open conformation is achieved by the end of the simulation. Even though there is no significant coupling between the flaps and saquinavir ($Cc(|\mathbf{r}_{FA}|:|\mathbf{r}_{SA}|) = 0.13$), there is a significant reduction in the coupling between the aspartic acid dyad and the drug ($Cc(|\mathbf{r}_{FA}|:|\mathbf{r}_{FS}|) = 0.25$), as compared to wildtype. This facilitates significant lateral motion, as indicated by the convergence of the Flap-Asp (black line) and Saq-Asp (orange line) distances. The G48V/L90M double mutant shows stable positioning of the drug

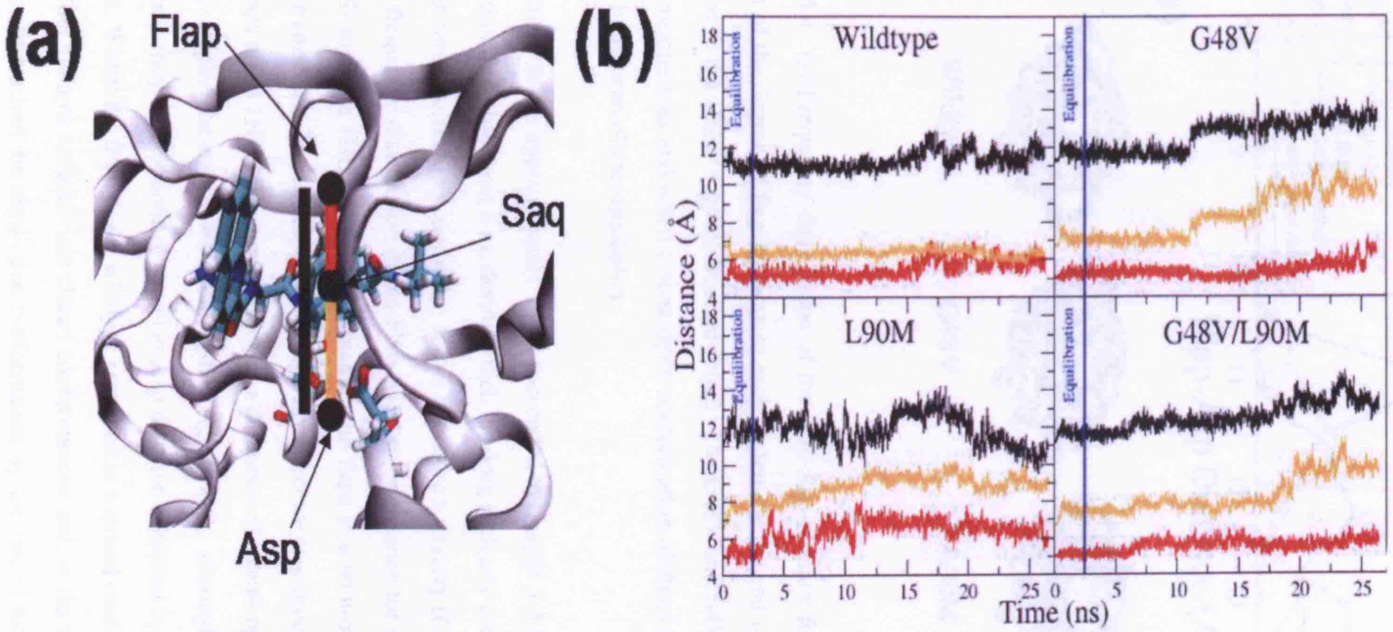


Figure 5.3: (a) Schematic diagram of saquinavir bound to HIV-1 protease. Three coordinates are shown (black circles), Flap, Saq and Asp, defined as the centres of mass of residues 49 of each monomer, all saquinavir atoms, and enzymatic aspartic acid C_β atoms respectively. The black, red and orange lines are the magnitudes of Flap-Asp ($|\mathbf{r}_{FA}|$), Flap-Saq ($|\mathbf{r}_{FS}|$) and Saq-Asp ($|\mathbf{r}_{SA}|$) vectors respectively. (b) Time evolution of the Flap-Asp, Flap-Saq and Saq-Asp vectors over 25 ns for a single representative trajectory of each protease system. Coupled flap-inhibitor motion as well as the transition of the flaps from a closed to a semi-open conformation are observed in the G48V and G48V/L90M systems.

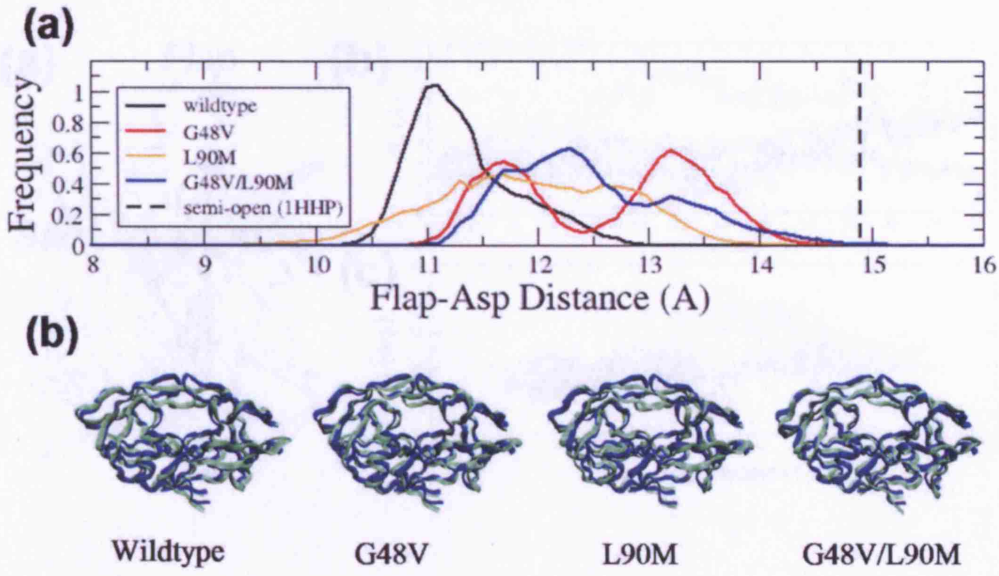
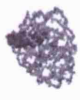


Figure 5.4: (a) Frequency distribution of the Flap-Asp distance for all four systems. (b) Schematic diagram of the degree of flap opening in each system by the end of the 25 ns simulations (green), as compared to the 1HHP crystal structure (blue). The G48V and G48V/L90M systems sample the most open conformations and reach a semi-open conformation (defined by the 1HHP Flap-Asp distance = 14.9 Å) by the end of the simulation.

in the active site for approximately 19 ns, succeeded by a rapid 3 Å motion away from the aspartic acid dyad within the subsequent 6 ns, during which the drug is clearly coupled to the flaps. This is supported by a high cross-correlation coefficient across the entire trajectory ($Cc(|\mathbf{r}_{FA}|:|\mathbf{r}_{SA}|) = 0.90$).

The frequency distribution of the Flap-Asp vector distance for all systems is shown in Figure 5.4. The PDB structure 1HHP of the apo protease with flaps in a semi-open conformation was taken as our reference structure. The corresponding Flap-Asp vector for this structure is 14.90 Å. If the conformation of the flaps in the 1HHP structure is taken as a definition of a semi-open conformation, then none of our systems sampled the semi-open conformation significantly, although there were significant differences in the frequency distributions of the Flap-Asp distance exhibited by the mutants as compared with the wildtype. Whilst the flaps of the wildtype remained in a closed conformation throughout the 25 ns simulation, the mutants sampled less closed conformations and, in the case of the G48V and G48V/L90M systems, approached the semi-open conformation by the end of the simulation (Figure 5.4). Furthermore, although an interesting bimodal distribution is discernible for the G48V mutant and to a lesser extent for the G48V/L90M mutant, as only one transition of the flaps from a closed to a semi-open conformation is observed in each of these systems, the aforementioned distribution does not represent



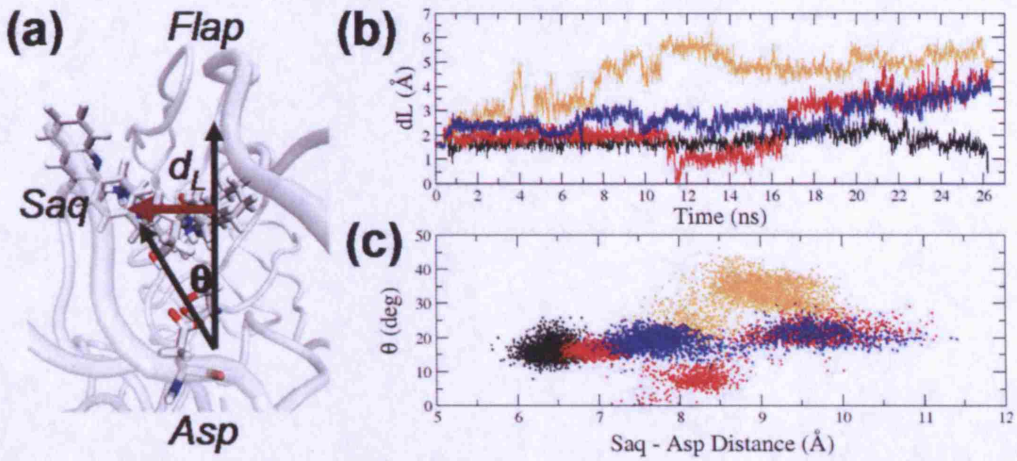


Figure 5.5: (a) Lateral motion of saquinavir out of the active site. The schematic diagram shows the lateral vector d_L measured by the perpendicular distance of the centre of mass of saquinavir (Saq) from the Flap-Asp vector (see Figure 5.3) and the angle θ between the Saq-Asp vector and the Flap-Asp vectors. (b) The time evolution of d_L and (c) the correlation of θ with the magnitude of the Saq-Asp vector are shown.

statistical conformational sampling.

5.3.3 Inhibitor Protrusion and Conformational Changes

Complementary to an analysis of drug motion along the active site axis, angular deviation of the Saq-Asp vector from the Flap-Saq vector together with the perpendicular distance of Saq from the Flap-Asp vector provide a direct indicator of lateral motion (Figure 5.5). The greatest lateral (d_L) and angular motion (θ) of the drug out of the active site are exhibited by the L90M mutant (orange line), which reaches a distance of 6 Å out of the active site only 10 ns into the simulation. G48V (red line) and G48V/L90M (blue line) mutants show trimodal and bimodal distributions respectively and in both the drug proceeds laterally to a perpendicular distance of 4 Å away from the Flap-Asp vector. Interestingly, in the G48V mutant, the drug first moves into the centre of the active site before being laterally directed away from the centre. Only in the wildtype (black line) does the drug maintain lateral proximity (perpendicular distance 1 Å to 2.5 Å) to the active site centre throughout the simulation.

The protrusion of the quinoline moiety of the P3 subsite (see Figure 5.6(a)) is observable from the alteration of the radial distribution function of water molecules around it at the end of the simulation as compared to the start. Figure 5.6(b) shows the radial distribution function of water molecules out to 10 Å from the C35 atom which is the quinoline atom furthest from the central plane of the active

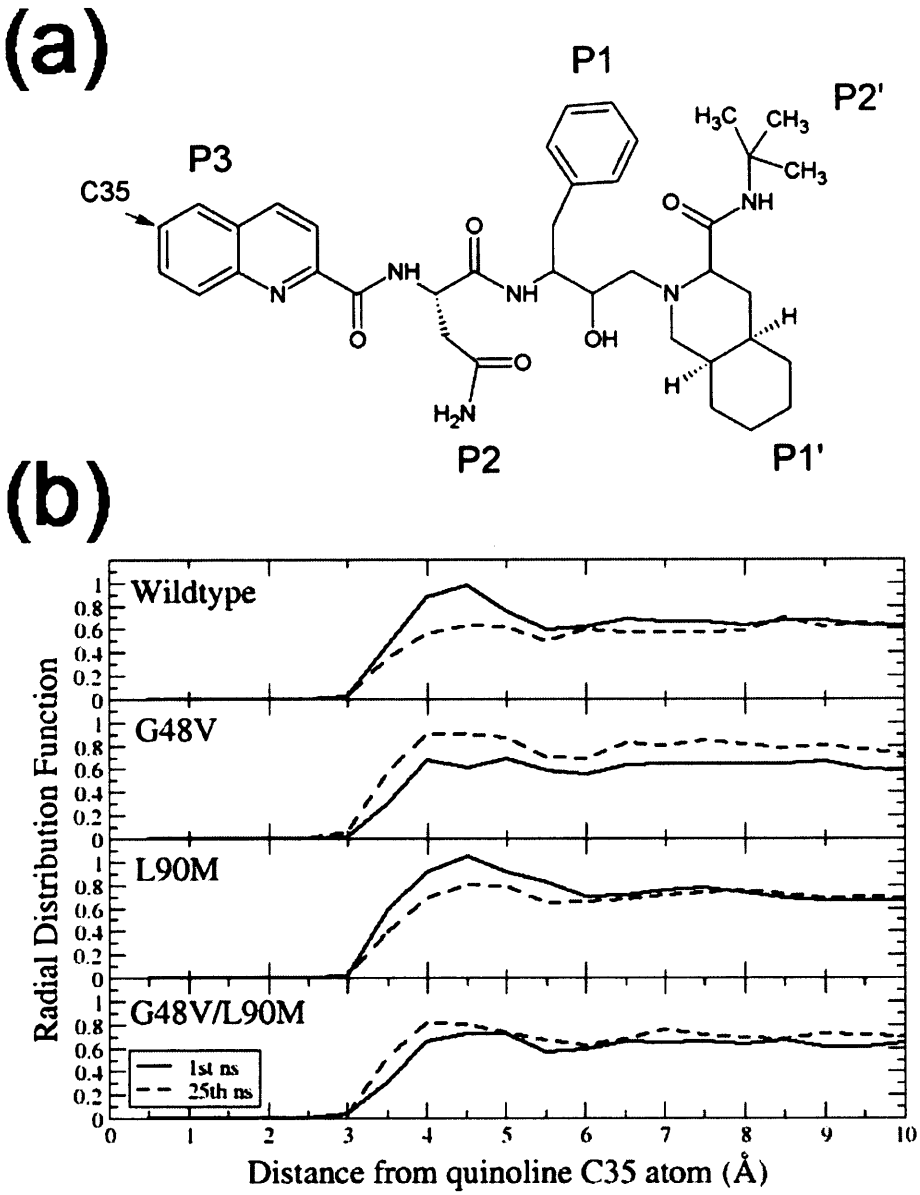


Figure 5.6: (a) Chemical structure of saquinavir showing subsites and the C35 atom on the quinoline moiety. (b) Radial distribution function out to 10 Å for all systems for both the 1st ns of production (solid line) as compared to the 25th ns (dashed line) of production from the C35 atom. There is increased solvation around the quinoline moiety in the G48V containing systems, associated with a pronounced protrusion from the active site.

site, averaged over 1 ns at the start of the production runs and at the end. The wildtype and L90M systems both show increased exposure to water at the start of the simulation as compared to the end, supporting an increased burying of the quinoline moiety into the active site. The G48V and G48V/L90M mutations show converse behaviour, in that the quinoline moiety is more exposed to water by the end of the simulation. In fact, for the G48V mutant, the extent of quinoline protrusion is large enough to cause significant increase in the radial distribution function even as far out as 10 Å from the C35 atom.

Conformational changes of both the drug and the protease were analysed further using both root mean squared deviation (RMSD) analysis and principal component analysis (PCA). RMSDs of the drug relative to its crystal structure were measured for two different alignment protocols (Figure 5.7(a)). In the first protocol (R_t), prior to calculation of the RMSD at each step, the protein backbones of the two relevant systems were aligned. In the second protocol (R_{cc}), the heavy atoms of saquinavir in both systems were aligned to each other. R_t therefore measures the total motion of saquinavir from its original position, incorporating bulk translational and rotational motion of the molecule as well as conformational changes, whilst R_{cc} measures solely conformational change. The difference between R_t and R_{cc} is an indication of the degree of bulk translational and rotational motion which we term the 'bulk' motion.

By the end of the simulation, R_t (dotted line) is larger in all mutants as compared to the wildtype. The largest R_{cc} (solid line) is observed for the L90M system, whilst saquinavir changes conformation little in the G48V and G48V/L90M systems. However, whilst this observation corresponds to lateral motion away from the active site centre in the L90M system, the change in R_{cc} in the wildtype corresponds to motion towards the catalytic centre (see Figure 5.5). In the wildtype R_t and R_{cc} are within 1 Å of each other at the end of the simulation implying little 'bulk' motion of the molecule. The increased separation of R_t from R_{cc} for all mutant systems demonstrates significant 'bulk' motion of saquinavir within these systems from its initial position and up to a 3 Å deviation from the active site centre by the end of the simulations.

Solvent accessible surface area plots of the averaged structure across the last 1 ns of simulation show elongation of saquinavir in the G48V and G48V/L90M mutants through rotation of the quinoline and phenyl moieties into a conformation where their planes are parallel to each other, and demonstrate the differential position of the drug in each of the four systems compared to the starting structure (Figure 5.7(b)). The larger degree of 'bulk' motion of all mutant systems compared with the wildtype together with the conformational changes associated with the wildtype are also confirmed by direct visual inspection.

Protease backbone atoms and all non-hydrogen drug atoms were used for PCA and all production trajectories were combined in order to produce a common set of principal component eigenvectors for all protease systems. The projections of the first three principal components for each system are shown in Figure 5.8. The zero-projection structures are shown in red and blue for the drug and protease

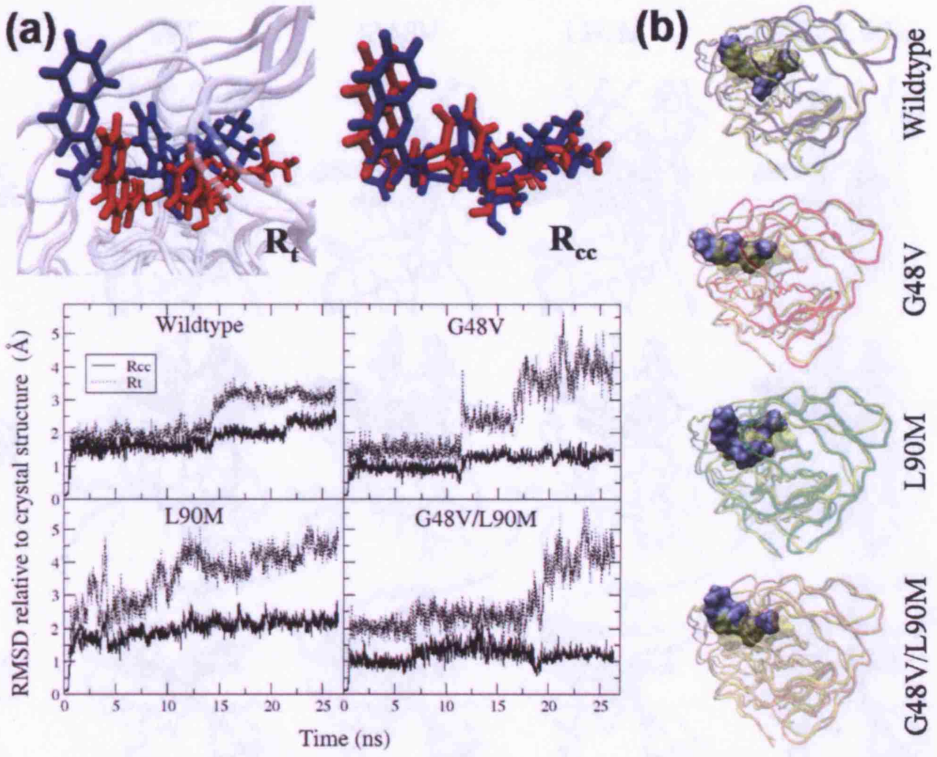
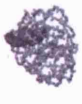


Figure 5.7: (a) Differences in the RMSD of saquinavir relative to its crystal structure in all four systems, for two different alignment protocols. R_t (dotted line) shows the RMSD of saquinavir atoms after alignment of the protein backbones, R_{cc} (solid line) after alignment to the heavy atoms of saquinavir. (b) Comparison of solvent accessible surface area plot (probe radius 1.4 Å) of saquinavir from the wildtype crystal structure 1HXB (transparent yellow) with the averaged structure over the last 1 ns of simulation (opaque ice-blue) for all systems. The backbones of the proteases are depicted in black, red, green and orange for wildtype, G48V, L90M and G48V/L90M mutant systems respectively and are aligned (R_t) before comparing the structures of the inhibitor.



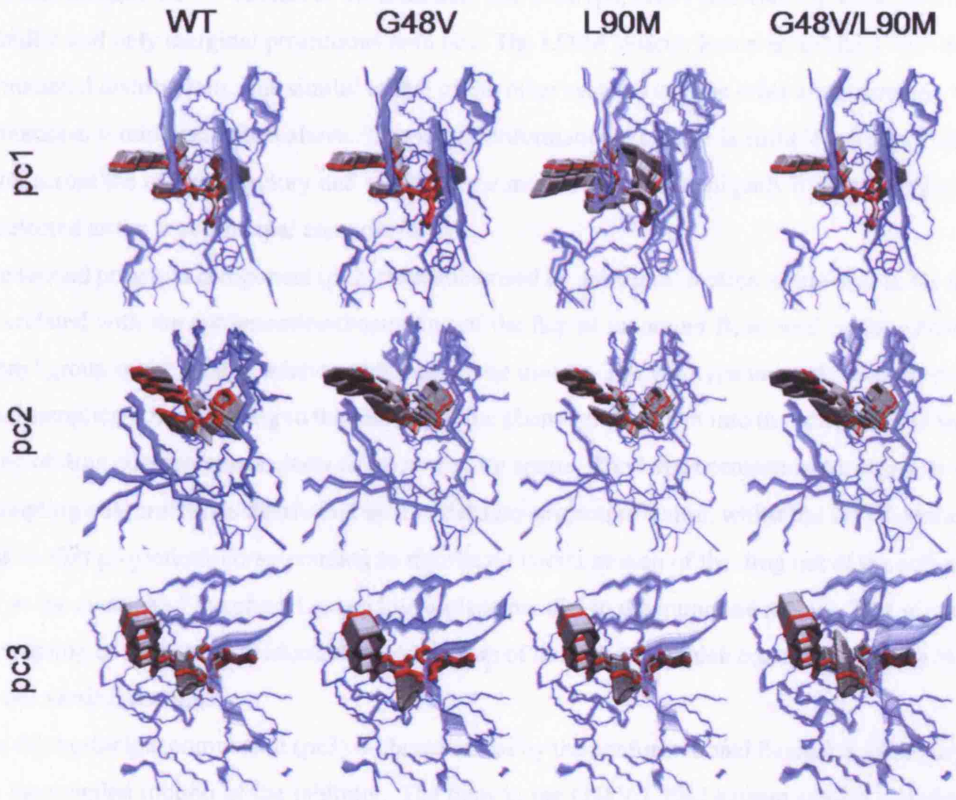
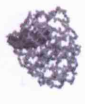


Figure 5.8: Principal component analysis of conformational sampling in each system. Superposition of the projections of each trajectory on the first three common principal component eigenvectors (pc1, pc2 and pc3) for all protease systems. The zero-projection structure is shown in red and blue for the drug and protease respectively and projections are shown in grey and ice-blue respectively. pc1 is dominated by conformational changes in L90M; the G48V-containing systems exhibit similar conformational sampling in pc2, the drug being laterally displaced from the active site centre with pronounced coupling of the drug to the flaps in both pc2 and pc3.



respectively and the corresponding projections onto each principal component, in grey and ice-blue respectively.

As viewed from the top of the flaps looking down, the first principal component eigenvector (pc1) is characterised by the correlated rotations of the drug, such that when the quinoline moiety rotates downwards, the P1′ subsite rotates upwards into the flaps. This correlates with the up-curling of the flap tips to accommodate the P1′ subsite of the inhibitor. The wildtype, G48V and G48V/L90M systems exhibit similar and only marginal projections onto pc1. The L90M system, however, exhibits two distinct conformational distributions, one similar to that of the other systems and the other corresponding to the large rotational motion described above. This large conformational change is sufficient to dominate the variation across the entire trajectory and results in the motion described uniquely by the L90M system being selected as the first principal component.

The second principal component (pc2) is characterised by the lateral motion of the drug in the active site, correlated with the conformational sampling of the flap of monomer B, as well as the rotation of the phenyl group of the drug in relation to the quinoline moiety. The wildtype uniquely exhibits conformational sampling corresponding to the rotation of the phenyl group down into the active site, as well as sampling of drug conformations closer to the active site centre. The L90M projection corresponds to the drug sampling conformations distributed around the zero-projection centre, whilst the G48V-containing systems exhibit projections corresponding to significant lateral motion of the drug out of the active site, as well as the rotation of the phenyl group into a plane parallel to the quinoline moiety. This is coupled to the sampling of more open conformations of the flap of monomer B, which contains a mixture of both lateral and vertical motions.

The third principal component (pc3) is characterised by the conformational flexibility of the flaps as well as the coupled motion of the inhibitor. The flaps in the G48V/L90M system sample significantly more of the conformational space than the wildtype, G48V and L90M systems. The G48V-containing systems exhibit the largest conformational sampling of the drug in relation to conformational changes in the flaps, indicating a more strongly coupled interaction between flaps and inhibitor as compared to the other two systems.

5.3.4 Differential Interactions in the Active Site

In order to investigate the molecular basis for the differential dynamics between the inhibitor and the various protease systems, especially the differential coupling to specific regions of the active site, we analysed both the hydrophilic and hydrophobic interactions within the active site across the course of the simulations.

Similar to the methodology adopted in Chapter 4, the active site of the protease is first decomposed into separate spatial sub-regions (see Figure 4.13(a) and 4.13(b)) in a way that encompasses all hydrophilic and hydrophobic interactions between the drug and the active site. The observed motion of

the mutant systems away from the centre towards the quinoline moiety 'exit' of the active site allows us to define an asymmetrical direction to the active site sub-regions. The components of the active site that are on the P3 - P1 side of the drug are therefore represented as 'front' subsites and those at the other end (P1′ and P2′) are represented as 'back'.

The 'Flap' sub-region consists of residues I47 to I50 (front) and I147 to I150 (back) that make up the inner strands of the flaps as well as the tetrahedrally co-ordinated water molecule between the flaps and the inhibitor, the 'Asp' sub-region is composed of the catalytic aspartic acid dyad and as it is central in the active site, it is not further designated as 'front' or 'back'. The 'Outer' sub-region consists of residues G27 to G30, R108 and L123 (back) and G127 to G130, R8 and L23 (front). Finally the 'Wall' sub-region of the active site consists of V32, P81, V82 and I84 at the front and V132, P181, V182 and I184 at the back. The 'Flap' and 'Outer' sub-regions therefore contain a mixture of hydrophilic and hydrophobic residues, whilst the 'Asp' sub-region is hydrophilic and the 'Wall' sub-region entirely hydrophobic. By comparison, the subsites of the drug are largely hydrophobic (P3, P1, P1′ and P2′); the only explicitly hydrophilic subsite is P2, identical to an asparagine residue. Other polar atoms are largely contained within the backbone of the inhibitor.

Hydrophilic interactions are assessed by calculating the running average number of hydrogen bonds, using a 100 ps time window, between saquinavir and all of the distinct sub-regions across the simulation span for each system. A donor-acceptor distance of 3.5 Å and a donor-hydrogen-acceptor angle of 150° were used as the criteria for the formation of a hydrogen bond (see Figure 5.9(a)). The total number of hydrogen bonds (black line) throughout the simulations is largest for the wildtype. Furthermore, in the wildtype, there is an isotropic distribution of hydrogen bonds between the 'Flap' (red line) and 'Asp' (blue line) sub-regions with running averages of 2.2 and 2.7 bonds respectively as well as almost no hydrogen bonding with the 'Outer' (orange line) sub-region. The wildtype also exhibits greater hydrogen bonding with the catalytic dyad than any of the mutant systems. This is explained due to the interactions of the P2 subsite of saquinavir within the active site.

In Chapter 4, we demonstrated that several distinct conformations of the P2 subsite of saquinavir can exist in each of the protease-inhibitor complexes considered here. These conformers are characterised by mutually exclusive hydrogen bonds and distinct dihedral angle distributions across the asparagine side-chain of the subsite (see Figure 4.7) and lead to differential coupling of the inhibitor with different regions of the active site. In conformation C_β, the P2 subsite is hydrogen bonded to a carbonyl oxygen of one of the catalytic aspartic acids and thus to the 'Asp' sub-region; in C_δ, the P2 subsite is hydrogen bonded to the backbone carbonyl oxygen of residue 148 in the flaps and thus the 'Flap' sub-region; in C_ε, there is a hydrogen bond with the backbone carbonyl oxygen of residue 127 and thus the 'Outer' sub-region. During these simulations, the P2 subsite evolves into conformation C_β for the wildtype which persists for the entire duration of the simulation. The P2 subsite evolves into C_δ for the G48V and L90M mutants for 17.9 ns and 10.8 ns respectively and firstly into C_ε for the G48V/L90M double

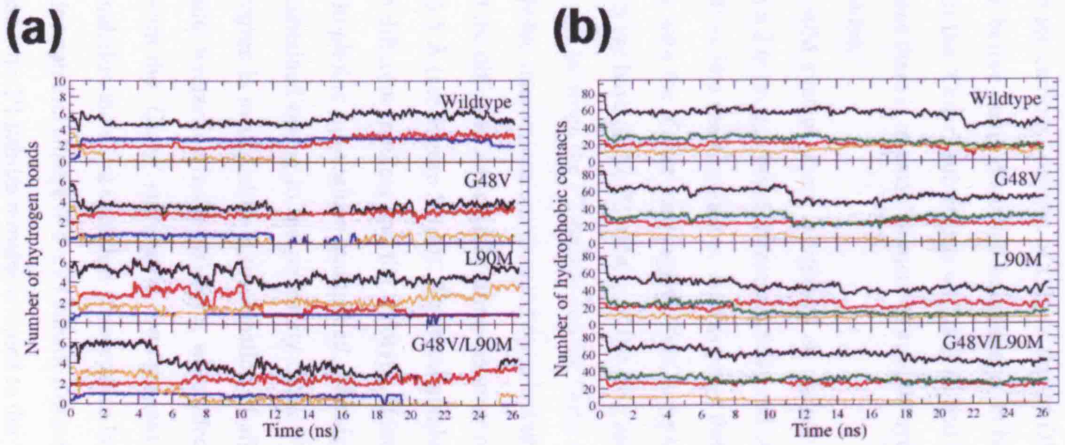


Figure 5.9: Time evolution of decomposed active site hydrophilic and hydrophobic interactions based on the sub-region decomposition shown in Figure 4.12. (a) Time evolution of inhibitor hydrogen bonds with the Flap (red), Asp (blue) and Outer (orange) sub-regions for each system (note the Wall (green) sub-region does not mediate any hydrogen bonds with the drug), using a 100 ps running average window. The wildtype maintains isotropy between the Flap and Asp sub-regions whilst in all other systems, Asp hydrogen bonding is significantly reduced and hydrophilic coupling to the flaps is larger. In G48V-containing systems, there is complete loss of Asp hydrogen bonds. (b) Time evolution of hydrophobic contacts with each sub-region (note the Asp sub-region does not mediate hydrophobic interactions). There is again a distinct loss of inhibitor contact with the Outer sub-region in G48V-containing systems. Interactions with all sub-regions are colour coordinated with Figure 4.13, the total interaction being represented by the black line.

mutant for a duration of 5.2 ns, followed by a transition into the C_δ conformation.

Due to this differential coupling in the G48V and G48V/L90M mutants as compared with the wildtype, there is at least one hydrogen bond more with the 'Flap' sub-region than with the 'Asp' for the entire duration of the simulation. The anisotropy is further increased after 12 ns and 19 ns respectively for these two mutant systems, at which point all hydrogen bonding with the 'Asp' sub-region terminates. The increased time to complete decay of hydrogen bonding between the inhibitor and the 'Asp' sub-region in the G48V mutant is due to the hydroxyethylene group of the inhibitor which switches acceptors from the carbonyl oxygen of D25 to that of D125, hydrogen bonding intermittently for several nanoseconds before being disrupted completely. In both these mutants, there is a decay of hydrogen bonding with the 'Outer' sub-region within 2 ns and 6 ns respectively. However, towards the end of the simulations there is re-establishment of a single hydrogen bond between the inhibitor and the highly polar R8 residue.

In the L90M system, there is initially an anisotropic distribution between 'Flap' and 'Asp' sub-regions with a 2 hydrogen bond difference, followed by isotropy between all three sub-regions between 11 ns and 20 ns into the simulation. At this point there is a marked increase in hydrogen bonding by 3 or 4 bonds with the 'Outer' sub-region. This is explained by the hydrogen bond established between the oxygen at the base of the P3 subsite of the drug and the R8 residue as well as the formation of two hydrogen bonds between the P2 subsite and the Outer sub-region.

Hydrophobic interactions are assessed by calculating the number of hydrophobic contacts between the drug and the different sub-regions using the same running average criteria and with the same cut-off distance of 3.5 Å (see Figure 5.9(b)). The total number of hydrophobic contacts does not vary largely between the different mutants and the wildtype. However, only in the wildtype is isotropy between all three hydrophobic sub-regions maintained. The significant number of contacts with the 'Wall' sub-region is maintained owing to the proximity of the P1, P1′ and P2′ subsites, whilst contact with the 'Flap' sub-region is mediated by a combination of all the hydrophobic subsites. Interestingly, unlike the hydrophilic interaction in the wildtype, which decays to zero, there is a pronounced hydrophobic interaction with the 'Outer' sub-region, owing again to the P1′ and P2′ subsites. At 15 ns into the wildtype simulation there is a complete convergence between the number of 'Flap' and 'Outer' contacts due to a conformational change in the P1 subsite towards L23. This again is unique to the wildtype; in all mutant systems, the P1 subsite remains coupled to the 'Flap' sub-region. In the G48V and G48V/L90M systems, there is a decay in the number of contacts with the 'Outer' sub-region at 12 ns and 19 ns which again coincides with the previously mentioned rise in the position of the drug with respect to the base of the active site and the protrusion of the inhibitor. Finally there is an inversion in the number of contacts with the 'Flap' and 'Wall' sub-regions in the L90M system, owing to a conformational change in the P1′ subsite by which it associates with the 'Flap' sub-region instead of the 'Wall'.

The running average number of water molecules within 3 Å around the catalytic dyad was calculated

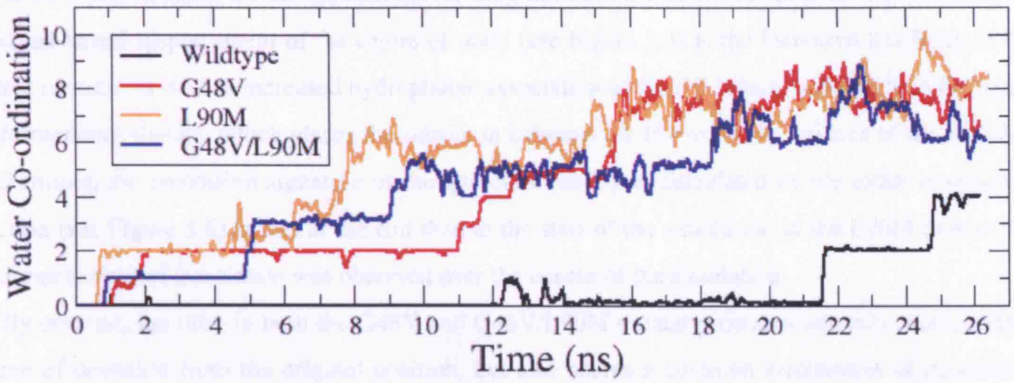
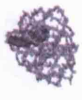


Figure 5.10: Water coordination within a 3 Å radius of the catalytic dyad, using a 100 ps running average window. Water ingress into this region occurs more for all mutants than for the wildtype. Furthermore, loss of hydrogen bonds with the aspartic acid dyad coincides with further water ingress in the G48V and G48V/L90M systems.

across the entire trajectory for each system (see Figure 5.10), again with a time window of 100 ps. In all mutants there are at least two water molecules coordinated around the catalytic dyad within 5 ns of the beginning of the simulation. In contrast, the wildtype protease exhibits no water coordination around the dyad until after 20 ns of simulation. The difference in water coordination between the wildtype and mutant systems ranges from 0 to 8 and by the end of the simulation there is still a coordination difference of 2 between wildtype and mutant systems. The previously mentioned sharp changes in the Flap-Asp and Saq-Asp vectors for both the G48V and G48V/L90M systems after 12 ns and 19 ns respectively, as well as termination of hydrogen bonds with the 'Asp' sub-region, coincide with increased water coordination around the catalytic dyad. In particular, at these points a water molecule enters into the cavity between the dianionic dyad and the hydroxyethylene group of saquinavir and in both systems leads to disruption of the hydrogen bond between the dyad and the inhibitor. For the wildtype and L90M systems, even though increased water coordination occurs, direct hydrogen bonding between the inhibitor and the dyad is not disrupted by any specific water molecule.

5.3.5 Mutation-Assisted Lateral Inhibitor Escape

To understand the molecular mechanism for the first stages of translation out of the active site, observed in all three mutants, and to assess the subsequent ease of lateral escape, we further investigated the change in the subsite interaction properties of the drug with various active site sub-regions across the course of the simulations, as well as laterally extracting the drug from the active site through steered molecular dynamics (SMD) simulations from the final conformation attained by the drug in each system.



In the L90M mutant, a clear mechanism for drug translation was not discernible. The cause of the observed lateral displacement of the centre of mass (see Figure 5.5) is the increased flexibility of the mutant protease as well as increased hydrophobic association of the P1′ subsite with the flaps through a conformational change, which places the subsite in between the I50 and I150 residues of the protease. Furthermore, the protrusion signature of the quinoline moiety as calculated by the radial distribution function (see Figure 5.6) is less at the end than at the start of the simulation in the L90M mutant: no clear mechanism of translation was observed over the course of the simulation.

By contrast, the drug in both the G48V and G48V/L90M mutant proteases not only shows a large degree of deviation from the original position, but also shares a common mechanism of translation. The properties of increased flap coupling combined with a loss of hydrophobic contacts and hydrogen bonds with the 'Outer' sub-region are common to both G48V containing mutants (see Figure 5.9(c) and 5.9(d)), as well as a similar signature of quinoline protrusion from the active site (Figure 5.6).

Figure 5.11 shows key drug-protease interactions in the G48V system at the start (Figures 5.11(a) and 5.11(b)) and the end (Figures 5.11(c) and 5.11(d)) of the simulation. At the start of the simulation, the P1′ and P2′ subsites form hydrophobic contacts with the back part of the 'Outer' sub-region (specifically with residues L23 and A28 respectively). These contacts are lost at 12 ns and 17 ns respectively (Figures 5.11e[1] and 5.11e[2]) and are correlated with an increase in hydrophobic contacts with the Flap and Wall sub-regions. This is due to a translation of the P1′ subsite away from L23 and a rotation of the P2′ subsite into the hydrophobic Wall cavity composed of V32, P81, V82 and I84 residues (Figure 5.11(c)). Furthermore, unlike the wildtype system in which the P1 subsite rotates to form hydrophobic contacts with the front 'Outer' sub-region, the presence of valine at position 148 allows maintenance of hydrophobic contacts with the flaps in the G48V containing mutants. This results in the P1 subsite, which initially also has hydrophobic contacts with the Wall cavity, moving clear of it by the end of the simulation. This is facilitated by the flexibility of the Wall tip, the Wall tip distance (P81-P181 C_α) varying by up to 5 Å across the simulation (Figure 5.11(e)[4]); such breathing allows the P1 subsite to move beyond the hydrophobic pocket initially constraining it.

As well as conformational changes in the drug, we noticed a large conformational change in the R8 residue which guards the exit to the active site (Figure 5.11(b) and (d)). At the beginning of the simulation, both 'gate' arginines R8 and R108 are approximately equidistant to the hydroxyethylene moiety of the inhibitor. Figure 5.11(e)[3] shows the evolution of these distances over the course of the simulation. At 17 ns into the simulation, there is a coupled movement of the OH group away from R108 (red line) and towards R8 (black line) as well as significant motion of R8 towards the OH group. This coincides with the loss of hydrogen bonding with the aspartic acid dyad (Figure 5.9(e)), following the flap induced lifting of the drug at 17 ns (Figure 5.3 orange line).

The R8 conformational change establishes a stable, strong hydrogen bond which lasts for the remainder of the simulation, preventing motion of the OH group back towards the catalytic dyad and

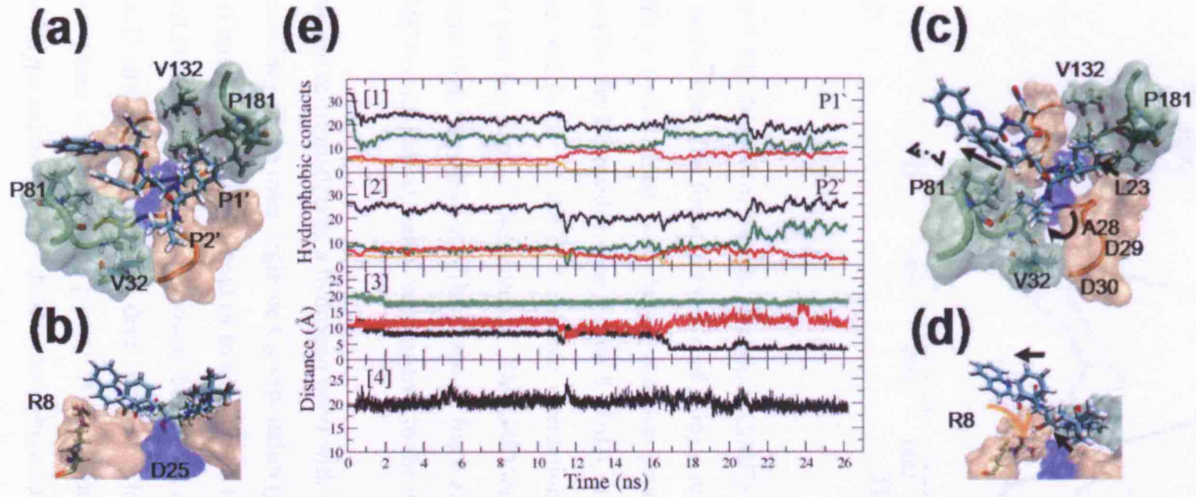


Figure 5.11: (a) Distinct inhibitor subsite contacts with active site sub-regions and (b) position of the R8 'gate' residue for the G48V mutant system at the start of the simulation. (c) Unidirectional translation of P1, P1′ and P2′ inhibitor subsites in the G48V system and (d) position of inward rotated R8 'gate' residue at the end of the simulation. (e) Time evolution of hydrophobic contacts of [1] the P1′ and [2] the P2′ subsites with the active site sub-regions, following the same colour scheme as Figure 5.9. Complete decay of contacts with the back Outer sub-region occurs as well as transfer of subsite contacts to the Flap and Wall sub-regions. [3] Interatomic distances for R8 and R108 C_α atoms (green line), saquinavir hydroxyethylene group oxygen SAQ:O - R8:NH1 atom (black line) and SAQ:O - R108:NH1 (red line). Whilst interatomic separation of the backbone of R8 and R108 gate residues remains constant, a combination of inhibitor motion towards R8 and away from R108 as well as rotation of R8 towards the hydroxyethylene oxygen results in the formation of a hydrogen bond. [4] Time evolution of the interatomic separation of P81 and P181 C_α Wall tip residues. The Wall tip breathes with distance fluctuations up to 5 Å, allowing for the passage of the P1 subsite past the front Wall sub-region in the course of the simulation.

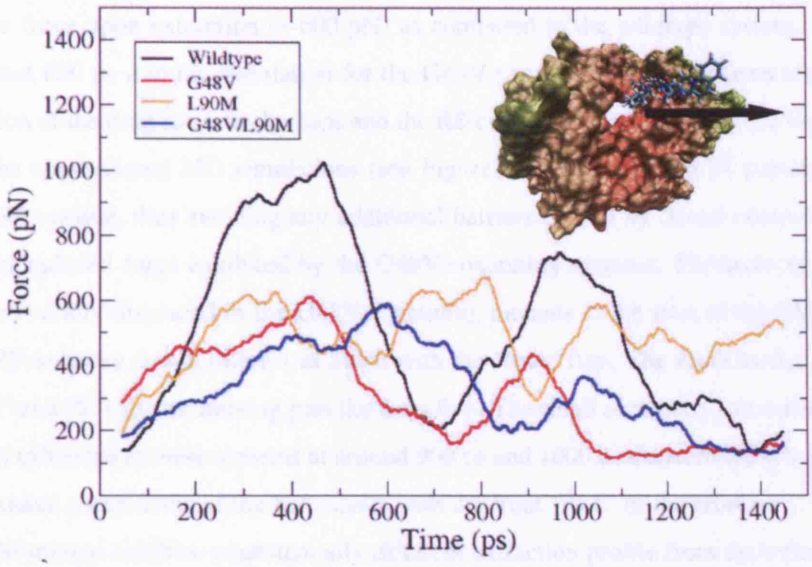
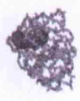


Figure 5.12: Steered molecular dynamics (SMD) lateral extraction of saquinavir from the HIV-1 protease active site. The force is calculated using a running average time window of 100 ps. The wildtype exhibits a well-defined two-phase extraction profile, the first barrier being due to the conformation adopted by the R8 residue (see Figure 5.11(d)), which obstructs the P1 subsite during drug extraction. The second is due to the hydrophobic interactions mediated by the P1′ and P2′ subsites as they are pulled past the flaps of the protease. The G48V-containing mutants are initially already more laterally displaced than the wildtype. The resistive force along the steered pathway is subsequently smaller for the G48V-containing systems as compared to the wildtype by approximately 400 pN.

subsequent re-formation of a hydrogen bond with it. Stability of the arginine backbone is confirmed by the constancy of the inter-arginine C_α separation (green line) and confirms a rotation of R8 towards the drug (Figure 5.11(e)[3]) instead of motion of the backbone. All of these mechanisms also occur in the G48V/L90M mutant (data not shown), but do not appear in either the L90M or the wildtype systems.

SMD simulations of lateral drug extraction from the active site reveal that the G48V-containing systems share a similar force profile to the wildtype but with varying magnitude (see Figure 5.12). The wildtype exhibits a well-defined two-phase extraction profile; the resistive force of the first barrier ($\sim$1000 pN) occurs at around 500 ps into the extraction and is due to the conformation adopted by the R8 residue (see Figure 5.11(d)), which obstructs the P1 subsite as well as the P1′ and P2′ subsites being pulled underneath the flaps. The second barrier ($\sim$750 pN) is due to the hydrophobic interactions mediated by the P1′ and P2′ subsites as they are pulled completely clear of the flaps of the protease approximately 1000 ps into the simulation.



The G48V-containing systems also exhibit a two-phase extraction profile with a significantly reduced peak resistive force upon extraction (~600 pN) as compared to the wildtype system. This occurs at around 440 and 600 ps into the simulation for the G48V and G48V/L90M systems respectively. The coupled motion of the drug towards the flaps and the R8-conformation adopted by the G48V-containing systems in the conventional MD simulations (see Figure 5.11(d)) allow the P1 subsite to be already clear of the R8 residue, thus avoiding any additional barriers caused by lateral obstruction of R8 and explaining the reduced force exhibited by the G48V-containing systems. Furthermore, as the drug is already more laterally displaced in the G48V-containing mutants at the start of the SMD simulations, the P1′ and P2′ subsites do not interact as much with the 'back' flap. The main barrier to overcome is that of the P1′ and P2′ subsites moving past the front flap. The small secondary force peaks of ~400 pN and ~350 pN exhibited in these systems at around 900 ps and 1000 ps respectively into the simulations are due to residual interactions of the P2′ subsite with the front 'Wall' of the protease.

The L90M mutant exhibits a substantially different extraction profile from the other three systems with a resistive force fluctuating around 500 pN for the whole simulation. This is due to the P1′ and P2′ subsites being caught between the flaps of the protease. Upon extraction the hydrophobic interactions between the front flap and the P1′ and P2′ subsites are maintained and the flap rotates laterally outwards as the drug is expelled.

Previous molecular simulations on the protease complexed with saquinavir have suggested a monoprotonated dyad with Asp 25 being thermodynamically favoured [242]. However, at physiological pH the catalytic dyad is dianionic and so the proton would have to bind after or upon ligand binding. We, therefore, also investigated the effects of altering the protonation state of the catalytic dyad. A full account of the study can be found in Chapter 6, where such simulations are used in the determination of equilibrium binding affinities, but we also provide a brief summary of the relevant dynamical results here.

Simulation of all four protease systems over a period of 10 ns, with the dyad in a monoprotonated state, reveal the comparative immobility of the drug and the flaps of the protease in these systems. The flaps remain in a closed conformation with a Flap-Asp distance of around 12 Å, whilst the drug shows no flap-coupling in any system, nor does it exhibit significant lateral motion out of the active site. Indeed, lateral motion is largest for the wildtype system, for which the centre of mass of the drug moves approximately only 1 Å from the starting position. The first stage of the G48V-assisted escape mechanism, reported here, is therefore facilitated more by a dianionic dyad state than a monoprotonated state and is thus able to take advantage of the dianionic state to confer drug resistance.

5.4 Discussion

The alteration in binding between an inhibitor and mutant forms of the HIV-1 protease compared to the wildtype has been well studied by a variety of methods and provides a quantitative basis for the assessment of drug resistance. However, measures of binding affinity alone cannot provide a causal mechanism of drug resistance at the molecular level.

Therefore, in order to provide molecular insight into the kinetic basis of drug resistance conferred by the characteristic G48V and L90M mutations of HIV-1 protease to the inhibitor saquinavir, and to observe the deviation of the drug from its initial crystallographic position, the differential dynamics of the two single mutants and the double mutant were compared to the wildtype protease bound to the drug, using molecular dynamics simulations in explicit water, over a timescale of 25 ns.

Our results confirm that the degree of isotropy of hydrogen bonding and hydrophobic contacts between the inhibitor and various sub-regions of the active site plays a key role in determining the subsequent dynamics of the inhibitor. Furthermore, this isotropy is in turn partially dependent on the conformation of the P2 subsite adopted by the drug. Only in the wildtype is an isotropic distribution between flap-inhibitor and catalytic dyad-inhibitor hydrogen bonds (Figure 5.9(c)) observed, and this is due to the adoption of the C_β conformation by the P2 subsite. Furthermore, hydrophobic isotropy is also maintained between the drug and different sub-regions of the active site (Figure 5.9(c)). Consequently, there is little bulk motion of the inhibitor within the active site across the entire 25 ns, as well as no change in the conformation of the flaps (Figures 5.3 and 5.4).

We observe a common mechanism of drug translation out of the active site in the G48V-containing mutants over the 25 ns period. The presence of the G48V in the protease not only increases hydrophobic 'Flap' interactions with the P3 and P1 subsites, but induces an extra hydrogen bond between the P2 subsite and the flaps. This is due to the adoption of conformation C_δ by the P2 subsite, although in the G48V/L90M mutant, such a conformation is preceded by C_ε. This together with the lack of significant hydrogen bonds with the catalytic dyad for these mutants has a pronounced effect (Figure 5.9).

Firstly, increased flap coupling facilitates motion of the inhibitor towards the flaps (Figure 5.3), leading to complete disruption of hydrogen bonding with the catalytic dyad (Figure 5.9) as well as coinciding with entry of water molecules into the catalytic region (see Figure 5.10). Secondly, it induces the flaps to sample more open conformations, in turn leading to further subsequent motion of the inhibitor away from the catalytic dyad.

Flap induced lifting of the drug away from the dyad also causes loss of hydrophobic interactions between the P1′ and P2′ subsites with L23 and A28 hydrophobic residues in the outer regions of the active site, resulting in the pronounced rotation of P2′ respectively into the Wall sub-region cavity composed of V32, P81, V82 and I84 (Figure 5.11). The P1 subsite moves clear of this hydrophobic well, facilitated by significant breathing of the Wall tip residues, and the quinoline moiety of the drug becomes more

exposed to solvent (Figure 5.6) due to a 4 Å lateral shift out of the active site (Figure 5.5). Furthermore, principal component analysis (PCA) confirms the significant increase in lateral conformational sampling by the drug in the G48V-containing mutants (see Figure 5.8(pc2)) as well as the coupled expulsive motion of the flaps. Such lateral drug translation induces a conformational change in the R8 active site gate residue, which subsequently forms a hydrogen bond with the emerging central hydroxyethylene moiety of the drug (Figure 5.11).

Although, by the end of the simulation, the drug has not been completely expelled from the active site, the distinctive mechanisms described above together with their occurrence in both G48V containing mutants provide compelling insight into a plausible mechanism of lateral drug escape from the active site. By the end of the simulation, the flaps in these mutants are in a semi-open conformation and the only remaining steric barrier to complete drug expulsion is the Wall tip residue P81 constraining the P2′ subsite. However, the breathing capability of this tip implies that over a longer timescale, such a steric barrier would diminish allowing for complete expulsion to occur. SMD simulations of the drug in the direction of lateral extraction (see Figure 5.12) confirm the relative ease with which the G48V-containing mutants can be extracted in comparison to the wildtype and are comparable with steered dissociation forces observed in studies of different biomolecular systems [58].

The increased sampling of more semi-open conformations of the flaps in the mutant proteases, especially the G48V and G48V/L90M systems, agrees well with previous computational studies on apo-proteases [114] over a similar timescale. In those studies the extent of flap opening is marginally more than in our studies owing to the absence of a bound inhibitor. Previous studies on the multi-nanosecond timescale have also shown that the flaps are stabilised in a closed position when bound to an inhibitor [142]. However, in those simulations, an implicit solvent was used resulting in decreased solvent viscosity. This results in enhanced flap opening in the apo-protease whilst the reverse effect is expected in the inhibitor bound case due to the increased apparent strength of flap-inhibitor interactions.

In our simulations, using explicit solvent, we show a transition to the semi-open conformation whilst the inhibitor is bound to the protease. The fact that G48V-containing systems reach a semi-open conformation by the end of the 25 ns simulation, in the presence of an inhibitor, is consistent both with an increase in sampling of semi-open conformations by mutant proteases, as well as a slightly reduced rate of opening of the flaps in an inhibitor-bound protease as compared to an apo-protease.

Furthermore, the significant deviation of the inhibitor away from the initial position reported here and the relative ease of extraction using SMD, together with the fact that full flap opening is rare compared to the semi-open conformation [142], provide a strong basis for the hypothesis that lateral drug expulsion from a semi-open protease is possible.

Binding of peptidomimetic inhibitors to the protease occurs by a previously cited two step process [255]:

$$E_o + I \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} E_o.I \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} E_c.I \tag{5.1}$$

where I represents the inhibitor, E_o.I and E_c.I represent 'loose' and 'tight' forms of the protease complex with the flaps in an 'open' and 'closed' form respectively and where k_1, k_{-1} and k_2, k_{-2} are the rate constants from the first and second steps of these complexes respectively. In such a process, the combination of the forward and backward rates give rise to observable association (k_{on}) and dissociation (k_{off}) rate constants from unbound to tightly bound proteases. It has recently been shown, using coarse-grained Brownian dynamics simulations, that binding of ligands is gated by the flaps, which modulate access to the active site [141]. Gating rates for the G48V and L90M mutations alone were comparable to the wildtype in these studies, whilst being smaller for proteases with several mutations including these two mutations. In the above context, changes in 'slow gating' observed in these studies correspond to alteration in k_1.

Previous authors attributed the decrease in binding affinity of saquinavir with mutant proteases, represented by an increase in k_{off}, to a decrease in the equilibrium constant ($K = k_2/k_{-2}$) between 'tight' and 'loose' forms of the complex [15]. Furthermore, complete dissociation of the inhibitor from the protease was only suggested from the flap-open protease.

Based on our simulations, which exhibit a tendency towards lateral expulsion, we propose a simultaneously occurring and alternative mechanism for the increase in k_{off} induced by the G48V mutant, which results in a decrease in the dissociation constant. In such a mechanism, complete dissociation can occur with the protease in a semi-open flap conformation:

$$E_{so} + I \underset{k_{-3}}{\overset{k_3}{\rightleftharpoons}} E_{so}.I \underset{k_{-4}}{\overset{k_4}{\rightleftharpoons}} E_c.I \tag{5.2}$$

E_{so} represents the unbound form of the enzyme with the flaps in a semi-open conformation. The rate constants k_3 and k_{-3} are for binding and unbinding of inhibitors from a semi-open apo-protease.

As binding of inhibitors is most likely diffusion limited [141], it is plausible that k_3 is sufficiently small as not to be an alternative viable mechanism for binding, thus preserving the original two-step binding process. From the bound state however, inhibitor dissociation may proceed via both mechanisms provided k_{-3} is not too small. Furthermore, as experimentally k_{on} is found not to change significantly in the G48V mutant, resistance may be manifested by an increase in k_{-3} as compared to the wildtype, as well as the increase in k_{-2} suggested by previous authors [15]. This slight increase would mean that it is additionally viable for the inhibitor to become unbound via a lateral expulsion mechanism before being re-stabilised in a flap-closed protease. In the case of the G48V mutant systems simulated here, it is the increased coupling of the flaps to the inhibitor that alters the position of the drug sufficiently explaining how k_{-3} could be increased whilst, even though some flap fluctuation is observed in the wildtype (Figure 5.3), the lack of extensive coupling prevents the drug from any large shift in

positioning. Furthermore, it is plausible that there may be a whole class of mutations for various inhibitors that take advantage of the increased rate of such lateral inhibitor expulsion to confer resistance to inhibitor binding. It would be very interesting to see the effects of such long timescale simulations on different inhibitor/mutant combinations.

Our study shows that the dianionic protonation state of the catalytic dyad plays a significant part in the assistance of the observed lateral motion. A mechanism of lateral dissociation in alternative protonation states has not been reported in the literature, nor in our own studies of the protease in the monoprotonated state, in which no lateral motion is observed over a timescale of 10 ns (see Chapter 6). As any protonation of the dyad most likely occurs upon or after ligand binding [156], our work shows that the G48V mutation is able to take advantage of the dianionic state of the protease to confer drug resistance through the expulsion mechanism described here.

The significant translation of the inhibitor in a concerted direction, provides the basis for an improved strategy in drug design. Steric hindrance of the large P3 subsite quinoline moiety with the flaps prevent lateral shift towards the P1′ and P2′ side exit of the active site in the semi-open conformation, thus acting as an 'anchor'. The lack of a sufficiently sized anchor at the other end of the drug, especially one that can oppose lateral motion through steric hindrance with the flaps, allows the drug to be translated unidirectionally. Inhibitors with symmetric and sufficiently sized anchors at both ends of the drug may enhance drug binding in the semi-open state and counter the translation effects induced by excessive flap coupling. Indeed, the lack of sufficient interaction with the catalytic dyad in the mutants studied here, promotes this excessive flap coupling. Inhibitors should therefore also be designed to incorporate stronger binding with the catalytic dyad.

CHAPTER 6

Quantitative Ranking of Drug Resistance of HIV-1 Protease Mutants Bound to Saquinavir using Free Energy Methods

FREE energy is arguably one of the most important physical quantities governing the interactions of biochemical systems. In the context of chemotherapeutic intervention, the strength of inhibitor binding is ultimately determined by the free energy difference of binding of a drug to its target. As such, the free energy difference of binding is a desirable quantity to ascertain in the quest to design effective inhibitors for a large array of pathogenic illnesses. In Chapter 1, we discussed several experimental methods by which the binding affinity of an inhibitor for a protein can be determined, whilst in Chapter 2 we discussed how it can be determined from an array of computational approaches.

With regard to the continuing problem of drug resistance in the treatment of HIV, discussed in Chapter 3, the change in the binding affinity of a set of inhibitors for drug-resistant mutational strains of retroviral enzymes such as HIV protease is an effective way of gauging the degree of resistance conferred by emergent mutations. In this chapter we explore the applicability of the MMPBSA method (described in detail in Chapter 2), which uses the framework of molecular simulation to calculate the free energy of binding of a ligand to a protein, to quantitatively rank characteristic drug resistant mutations of HIV-1 protease.

We also report the development of a tool for the automated calculation of binding affinities of HIV-1 protease complexes, termed the 'Binding Affinity Calculator' (BAC), which has been used to facilitate both the studies reported here and those reported in Chapter 7 and which can be used to rapidly construct models, implement simulations and calculate binding free energies of a large array of HIV-1 protease-ligand variants.

6.1 Background

Since the emergence of the first anti-retroviral inhibitors to treat HIV, extensive studies have been conducted both in the experimental and computational domain, in order to quantify the strength of binding of many inhibitors to various target proteins of the virus. Furthermore, this effort has included a substantial attempt to understand the continuing problem of drug resistance in terms of the reduction in the binding affinity of various inhibitors to the proteins of emergent mutant strains of the virus.

Experimentally, techniques such as enzyme inhibition assaying (EIA) and isothermal titration calorimetry (ITC), discussed in Chapter 1, have provided a wealth of data regarding both the HIV reverse transcriptase and protease enzymes [16, 17]. The EIA technique has been the most common method applied to determining reductions in binding affinity for a range of inhibitors to the HIV-1 protease wildtype and many mutants, where usually values of the inhibition constant K_i, as well as enzymatic parameters K_m and k_{cat} are obtained (see § 1.5.2). Also common are assessments of IC_{50} and IC_{90}, the concentrations of inhibitor required to reduce the enzymatic activity by 50% and 90% respectively (see § 1.5.3). One limitation of IC_{50} and IC_{90} assessments however, is that alone they only confer an indication of resistance and cannot differentiate between the cause of this resistance as being due to either a reduction in the binding affinity between the enzyme and the inhibitor or a reduction in the binding properties of the natural substrate. Determination of the inhibition constant K_i on the other hand is direct evidence of a reduction in binding affinity with the inhibitor.

Recent applications of the ITC technique have yielded much information regarding the binding of ligands to the protease [198, 201]. Furthermore, the use of the ITC method not only allows the free energy difference of binding to be measured, but furthermore provides a decomposition of the free energy into its enthalpic and entropic components. These studies have shown that several mutations reduce binding by altering either the entropic or enthalpic contributions to the free energy [197, 198].

The free energy difference of inhibitors binding to both wildtype and mutant HIV-1 proteases has also been studied using a range of computational approaches including linear response (LR) [258], MMPBSA [218, 245, 259, 260] and thermodynamic integration (TI) methods [261]. As discussed in Chapter 2, these methods vary in accuracy and computational expense [60, 262]. Whilst the LR approach is much faster, it relies on statistical regression using empirical data, unlike MMPBSA and TI. TI on the other hand, whilst the most accurate, is limited to providing only the relative free energy differences of binding between systems as well as being very computationally expensive.

In this regard, the MMPBSA approach is appealing, not only because it requires no experimental fitting [263] and is relatively inexpensive computationally compared to TI, but also because it can yield absolute free energy differences of binding (see Chapter 2). However, unlike TI which is thermodynamically 'exact' within the classical approximation, the approximations inherent to MMPBSA also result in larger errors than those associated with more accurate methods and potentially greater discrepancies

with experimental results [245]. Despite its limitations, the MMPBSA methodology has been applied, with varying degrees of success, to biomolecular systems as diverse as avidin [264], cathepsin D [265], matrix metalloproteinases [263], as well as HIV-1 reverse transcriptase [266].

An early MMPBSA study conducted on HIV-1 proteases complexed with a series of inhibitors [218] has provided insight into the free energy profile of various drugs and the susceptibility of various positions in the protein sequence to develop mutations. A more recent study has attempted to characterise the binding properties of several inhibitors [245]. However, MMPBSA considerations alone do not take into account the contribution to free energy from the change in configurational entropy upon binding. This is a non-negligible component of the free energy and its inclusion is important in improving the accuracy of calculations. Neither study provides accurate values for the absolute binding free energy. In the study by Wang and Kollman [218], this is largely due to the fact that the configurational entropy has been ignored. Recent studies have shown that for amprenavir bound to the HIV-1 protease, the configurational entropy can be as large as ~25 kcal/mol [267]. In the study by Lepsik *et al.* [245], whilst the entropic contribution is included, accurate absolute values are still not obtained. Furthermore, a quantitative recipe for correctly ranking the binding properties of drug resistant mutants of HIV-1 protease to a single inhibitor using these methods, has not yet been achieved.

In this study, we use molecular dynamics simulations in explicit water alongside the MMPBSA method including entropic considerations from normal mode analysis to determine the absolute free energies of binding of the inhibitor saquinavir to the wildtype, the G48V, L90M and G48V/L90M mutants of HIV-1 protease. Furthermore, we explore the applicability of the MMPBSA approach combined with analysis of the entropic contribution in determining absolute free energies of binding as well as predictively and accurately ranking drug resistant mutants of the HIV-1 protease through an alteration in the drug binding affinity.

6.2 Methods

The initial preparation and simulation protocols implemented on the saquinavir-bound HIV-1 proteases in this study were very similar to those presented in Chapters 4 and 5. We will describe the differences here as well as describing the methods implemented to determine the enthalpies and configurational entropies of binding.

6.2.1 Initial Preparation of Models

The 1HXB crystal structure was used as the starting point for the wildtype protease bound to saquinavir whilst the 1FB7 structure was used for the G48V, L90M and G48V/L90M mutants. Unlike 1FB7, the 1HXB crystal structure contains two resolved rotationally symmetric sets of coordinates for saquinavir. The first set of drug coordinates were extracted to build the wildtype model. Drug charges for all systems

were assigned as described in Chapter 4. The standard AMBER forcefield for bioorganic systems (ff03) [24] was used to describe the protein parameters.

The protonation state of the aspartic acid dyad was also considered. As the study presented here is concerned with equilibrium thermodynamic quantities such as the binding free energy it was important to assign the most thermodynamically favourable end state of the protease-drug complex. Previous molecular simulations on the protease complexed with saquinavir have suggested a monoprotonated dyad with Asp 25 being thermodynamically favoured [242]. Therefore, even though a kinetic mechanism of drug dissociation is facilitated by the dianionic state (see Chapter 5) prior to protonation, for binding affinity calculations it was desirable to model the catalytic dyad in this monoprotonated state. The free energy calculated in this study was compared to two binding affinity experiments ε_1 and [15] and ε_2 [194] conducted at pH 6.5 and pH 5.0 respectively. As at these values of pH the protease is monoprotonated (see Chapter 3); it was therefore additionally important to select this protonation state for our binding free energy calculations.

The Leap module [248] in the AMBER 9 software package [53] was then used to combine each apo-protease system with the ligand, whilst retaining the crystal structure water molecules. Five Cl^- counterions were added to electrically neutralise each inhibitor-bound system. Each system was then solvated using atomistic TIP3P water [249] in a cubic box with at least 14 Å distance around the complex. The size of each inhibitor-bound system was 45314, 42680, 42670 and 42676 atoms for the wildtype, G48V, L90M and G48V/L90M systems respectively.

6.2.2 Minimisation and Equilibration Protocols

The molecular dynamics package NAMD2 [34] was used throughout the production simulations as well as for the employment of minimisation and equilibration protocols. Minimisation was conducted using the conjugate gradient and line search algorithms available in NAMD2 for 2000 iterations for each system with a force constant of 4 kcal/mol/Å^2 applied to all restrained atoms. This achieved a desired gradient tolerance of approximately 10 eVA^{-1} in each case. Restrained atoms included all heavy atoms of HIV-1 protease and the ligand.

The long range Coulombic interaction was handled using the particle mesh Ewald summation method (PME) [250]. A non-bonded cut-off distance of 12 Å was used for all simulations. For the equilibration and subsequent production runs the SHAKE algorithm [38] was employed on all atoms covalently bonded to a hydrogen atom, allowing for an integration timestep of 2 fs.

Each system was gently annealed from 50K to 300K over a period of 50 ps. The systems were then maintained at a temperature of 300K using a Langevin thermostat with a coupling coefficient of 5 /ps for the rest of the equilibration and for all subsequent production runs. All subsequent stages were carried out in isothermal isobaric (NPT) ensemble using a Berendsen barostat [49] with a target pressure of 1 bar and a pressure coupling constant of 0.1 ps. The systems were equilibrated for 200 ps whilst

maintaining the force constants on the restrained atoms to allow for thorough solvation of the complex and to prevent premature flap collapse [251].

This was followed by a mutation relaxation protocol to allow optimal re-orientation of all mutated amino acids. The heavy atoms of each mutated amino acid and those of amino acids within a 5 Å surrounding region of the mutation were completely relaxed sequentially for every mutation for a duration of 50 ps each. After each mutant region was relaxed for 50 ps, the heavy atoms of that region were again constrained by a force of 4 kcal/mol/Å^2 before proceeding to the next step.

This procedure was followed by a gradual force reduction on the ligand from 4 - 0 kcal/mol/Å^2 over a 200 ps period in equal stages of 1 kcal/mol/Å^2 and then a similar force reduction on the protease from 4 - 1 kcal/mol/Å^2 over a period of 150 ps. In the final stage of the equilibration, all constraints were removed from the protease and the system was allowed to evolve completely unrestrained up to a total duration of 2 ns. The length of this last stage therefore varied only according to the number of mutations that were incorporated in the system.

6.2.3 Production Runs

The production simulations for each system lasted 10 ns and were also performed in the isothermal-isobaric ensemble described above. Coordinate trajectories were recorded every 1 ps throughout all equilibration and production runs.

6.2.4 MMPBSA Calculations

The basis of the MMPBSA methodology has been described in detail in Chapter 2. Here we will describe some of the specific parameters employed for the implementation of each of the components of the MMPBSA calculation in this study.

As discussed in Chapter 2, there are two approaches for generating the required coordinate trajectories for the complex, protein and ligand. Either separate simulations are conducted for each molecular species or a single simulation is conducted from which the coordinates of the different molecular species are extracted. In this study we used the latter approach. Each component of the free energy is calculated for a range of snapshots across the desired coordinate trajectory for each species. The free energy difference of each component of the calculation is given by Equations 2.55 and 2.56. The total free energy difference of binding from the MMPBSA calculation is composed of the following terms:

$$\Delta G_b^{MMPBSA} = \Delta G_{vdW}^{MM} + \Delta G_{ele}^{MM} + \Delta G_{pol}^{sol} + \Delta G_{nonpol}^{sol} \tag{6.1}$$

where the first two terms on the right hand side represent the van der Waals and electrostatic components of the gas-phase molecular mechanics, free energy difference respectively, the third term is the

electrostatic/polar component of the solvation free energy and the last term is the non-polar component of solvation free energy (see § 2.5.2).

The MMPBSA module in AMBER 9 was used to implement the calculation for each MMPBSA component. The average molecular mechanics energy ΔG^{MM} was calculated using the SANDER module in AMBER 9, with no cut-off for the non-bonded energies. The AMBER PBSA module was used for the evaluation of the electrostatic free energy of solvation ΔG^{sol}_{pol}. A grid spacing of 0.5 Å was employed for the cubic lattice, the internal and external dielectric constants were set to 1 and 80 respectively, and 1000 linear iterations were performed. The non-polar solvation free energy ΔG^{sol}_{nonpol} was calculated from the solvent accessible surface area (SASA) using the MSMS program [67], with a probe radius of 1.4 Å, the surface tension γ set to 0.00542 kcal/(mol$Å^2$) and the off-set β to 0.92 kcal/mol.

The mean of the binding free energies of all the snapshots (N) used was computed and the standard error (σ) of the calculation was determined from the standard deviation (σ_{sd}) of the data set, where $\sigma = \sigma_{sd}/N^{1/2}$.

A minimum time interval between successive snapshots was determined for the MMPBSA calculation by determining the autocorrelation function in the first nanosecond of the G48V production trajectory. The MMPBSA calculation was performed on all 1000 snapshots (every 1 ps) and the autocorrelation function $C_f(t)$ for each component (f) of the calculation was calculated relative to the mean value ($\langle f \rangle$) of the 1000 snapshots across the 1 ns trajectory:

$$C_f(t) = \frac{\langle (f(t_i) - \langle f \rangle).(f(t_i + t) - \langle f \rangle) \rangle}{\langle (f(t_i) - \langle f \rangle)^2 \rangle} \tag{6.2}$$

Multiple time origins (t_i) were used from the beginning of the data set to the midpoint. Correspondingly, it was possible to calculate the function up to a time interval (t) of 500 ps where the value at each time interval was computed from a data set of 500 values. All MMPBSA calculations performed after this were implemented by extracting 1000 equally spaced snapshots (1 every 10 ps) from the 10 ns trajectory of the production run for each system.

This number of snapshots and the time interval between them were found to produce stable convergent free energies, with little computational expense (see § 6.3). Higher numbers of snapshots were not found to yield a significant gain in accuracy, which is consistent with reports of enthalpy calculations in previous MMPBSA applications [260]. On the other hand, a smaller time spacing between the snapshots makes it problematic to compute the variance of free energy averages because of the persistence of motional correlations on such time scales [262].

6.2.5 Calculation of the Entropic Contributions

The changes in configurational entropy upon ligand association $T\Delta S$ were estimated by an all-atom normal mode analysis performed with the AMBER NMODE module. When applying a harmonic ap-

proximation for the calculation of configurational entropy, the accuracy of the method is very sensitive to transitions between conformational wells [262]. Unfortunately, the selected 10 nanosecond portions of the trajectories did exhibit some conformational transitions, although these were limited to the P2 subsite of the inhibitor. Care was therefore taken when interpreting the results. Prior to the normal mode calculations, the complex, receptor and ligand were subjected to minimisation with a distance-dependent dielectric constant $\varepsilon = 4r$ and convergence tolerance tighter than drms = 10^{-4} kcal/molÅ. Since individual MD snapshots may adopt different conformations after minimisation, leading to differences in entropic contributions up to 5 kcal/mol [268], we performed an extensive analysis of the entropy across the 10 ns for each system. Entropy calculations on all protease-ligand systems were averaged over 50 equally spaced snapshots, extracted over the entire 10 ns of the production phase. Additionally, the entropy over the first nanosecond of production was determined from 20 equally spaced snapshots. The absolute free energy difference of binding, ΔG_b was then computed using Equation 2.67.

6.2.6 Computational Requirements

The molecular dynamics simulations were performed under conditions of optimal computational efficiency with a wall-clock rate of approximately 6 hours/ns, using 32 processors on Lonestar at the Texas Advanced Compute Center. One MMPBSA free energy calculation (1000 snapshots) required $\sim$ 30 hours CPU time. As the implementation of normal mode analysis in the AMBER NMODE module is serial, entropy computations using a normal mode treatment were more expensive, and each calculation involving 50 snapshots required $\sim$150 hours on a single Opteron CPU.

6.2.7 Automation of Binding Affinity Calculations

Due to the large array of parameters that need to be configured in a physically representative molecular simulation (see Chapter 2), the building of computational models for molecular modelling can be both involved and time-consuming. Furthermore, workflows involved in the implementation of molecular simulations, starting from the minimisation and equilibration steps right through to production runs marshalled across various high performance computational resources, as well as post-production analyses of free energies of binding, can be rather lengthy and tedious. However, studies often require the construction of models that vary only slightly and furthermore, the implementation of molecular simulations in such studies often follow very similar protocols.

For example, in the study implemented in this chapter, the only difference in the construction between the models of the four varying protease complexes was the incorporation of different mutations on the respective starting crystal structures of each system as well as slightly different equilibration protocols. These varied in the relaxation of the constraints on the system around the mutations that were incorporated. All simulations needed to be marshalled around high performance resources for compu-

tation of the output data files, which were subsequently transferred back to a data storage resource for subsequent determination of the free energies of binding in post-production analysis. Furthermore, the free energy calculation for all of these four systems, which itself required the construction of additional configuration files and execution on computational resources, was identically implemented.

Given such similarity in the workflow required to implement free energy calculations of multiple HIV-1 protease complexes, we have developed a tool, called the 'Binding Affinity Calculator' (BAC), that can automate the protocols of model building, simulating and post-production free energy analysis for a host of protease-ligand variants. By decomposing the workflow of a free energy calculation into these three stages, the BAC takes advantage of the Application Hosting Environment (AHE), discussed in § 2.6, to sequentially execute each stage of a calculation across various computational resources. It is particularly designed to utilise grid resources available through the AHE interface, for the automated marshalling of computationally demanding and data intensive simulations. Alternatively, each individual component can be adapted and used separately for simulations that vary beyond the scope of the BAC's design.

The BAC has facilitated the studies implemented here as well as those that will be reported in Chapter 7, in which HIV-1 protease variants are modelled in complex with the NC-p1 substrate. Furthermore, the BAC has been designed to facilitate future molecular dynamical studies of many drug-bound and substrate-bound studies of HIV-1 proteases in a high-throughput manner. In principle, the BAC can be integrated into clinical decision support systems to provide an additional tool for the discrimination of the resistance conferred by HIV-1 protease mutants against a host of available protease inhibitors. The motivations for such a tool, as well as the workflow and architecture of the BAC, are described in more detail in Appendix A.

6.3 Results and Discussion

6.3.1 Structural and Dynamical Properties of Monoprotonated HIV-1 Protease/Saquinavir Complexes

As a precursor to quantitative analysis of the free energy of binding, we determined several structural and dynamical properties of the drug in the active site of the wildtype, G48V, L90M and G48V/L90M mutant monoprotonated HIV-1 proteases. Such analysis allowed insights into the structural stability of each system over the 10 ns trajectory and also enabled a comparison with the study of the dianionic protease systems presented in Chapter 5.

Protein backbone RMSDs were calculated for each system relative to their original crystal structures. All systems except the wildtype exhibited stable backbone RMSDs across the entire 10 ns trajectories, with mean RMSDs of 1.02 $\pm$ 0.08 Å, 0.97 $\pm$ 0.07 Å and 1.03 $\pm$ 0.10 Å for the G48V, L90M and

G48V/L90M systems respectively. There were marked changes of ~0.5 Å in the wildtype backbone RMSD for the drug between the second and the fourth nanoseconds inclusive, followed by stabilisation of the RMSD with a mean of 1.21 ± 0.09 Å.

We determined the same Flap-Asp (black line), Flap-Saq (red line) and Saq-Asp (orange line) distances as in the dianionic study described in Chapter 5 (see Figure 6.1). The Flap-Asp distance remained stable at around 12 Å for all systems. Furthermore, the G48V, L90M and G48V/L90M systems exhibited almost identical Flap-Saq and Saq-Asp distances across the whole trajectory, remaining at 5 Å and 8 Å respectively. The wildtype exhibited conformational readjustments for just under 6 ns into the production phase, followed by stabilisation of the Flap-Saq and Saq-Asp distances at approximately 6 Å and 8 Å respectively. Therefore, in comparison to the 14.9 Å defined as the semi-open flap conformation in the 1HHP crystal structure (see Chapter 5), all systems remained in the closed conformation for the whole duration of the simulation. Furthermore, there was no evidence of coupled motion between the Flap-Asp and either the Flap-Saq or Saq-Asp vectors; the magnitudes of cross-correlation coefficients for all distance pairs were all less than 0.7, and no G48V-related motion of the inhibitor towards the flaps was observed.

The lateral motion of the inhibitor was also determined by calculating the lateral distance vector (d_L), also described in Chapter 5, which represented the perpendicular distance between the Flap-Asp and Saq-Asp vectors. In each system, the drug exhibited little lateral motion away from its original position of d_L ~2 Å. Indeed the wildtype moved the most with a mean d_L of 2.68 ± 0.33 Å. The G48V, L90M and G48V/L90M mutants also exhibited stable fluctuations with means of 1.93 ± 0.21 Å, 1.78 ± 0.18 Å and 2.18 ± 0.27 Å respectively. This contrasted with the lateral motion observed in the dianionic study, in which lateral motion up to 4 Å was observed from the original position.

Conformational changes of the saquinavir molecule as compared to the net motion of the drug were also determined using RMSD analysis (see Figure 6.2) across two different alignment protocols, R_t (alignment of the protease backbone) and R_{cc} (alignment of the heavy atoms of saquinavir). The wildtype exhibited significant 'bulk' deviation in the first 6 ns post-equilibration, but subsequently stabilised with a total RMSD (R_t) of just under 3 Å, with almost no separation of R_{cc} from R_t. However, the sudden increase of R_{cc} just after 6 ns of simulation was indicative of a sharp conformational change at that point in the production run. All mutant systems exhibited no significant change in RMSD post-equilibration. Furthermore, there was very little separation of R_t from R_{cc} in these systems, confirming that little 'bulk' motion of the inhibitor occurred from its original position.

The increased initial flexibility of the wildtype system, as indicated by the 3 Å difference in R_t (see Figure 6.2) as well as the increased fluctuation of the Flap-Saq-Asp distances (see Figure 6.1), followed by its stabilisation after approximately 6 ns can be explained by the motion of the P3 subsite of the inhibitor. In the first 6 ns of production, the P3 subsite adopted similar conformations to that of the 1HXB crystal structure, associating predominantly with residues G48 and P181 of the S3 subsite,

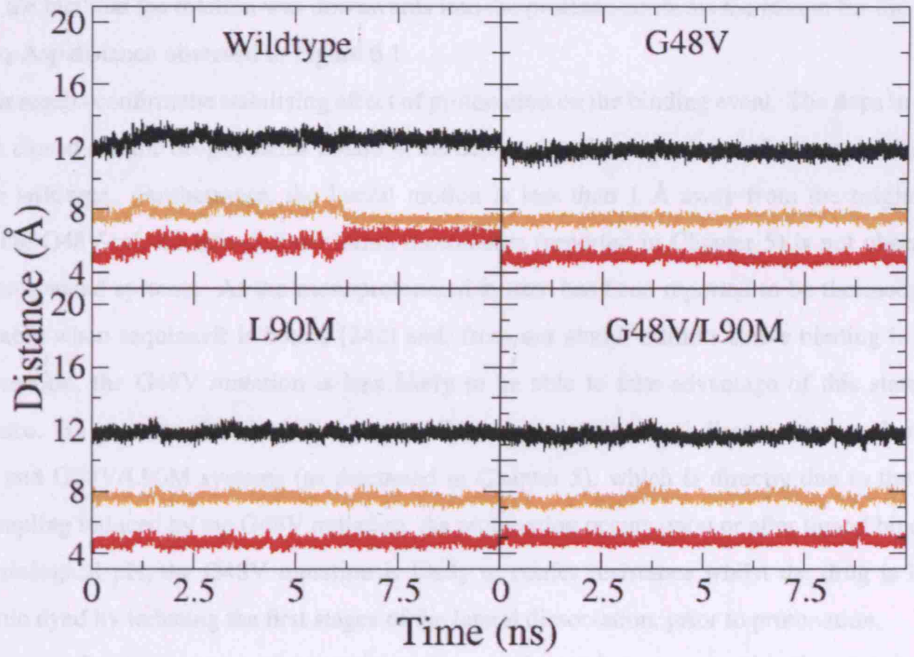
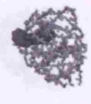


Figure 6.1: Time evolution of the Flap-Asp (black), Flap-Saq (red) and Saq-Asp (orange) vectors over 10 ns for each monoprotonated protease system. Unlike the dianionic proteases, the flaps remain in a closed conformation with a Flap-Asp distance of approximately 12 Å. Furthermore, there is no consistent G48V-related motion of the inhibitor towards the flaps.



residue V182 of the S1 subsite and with the quinoline moiety being parallel to the phenyl moiety of subsite P1. However, just after 6 ns there was a sharp change in the conformation of the P3 subsite into a conformation perpendicular to the plane of the P1 phenyl moiety, formed principally through an approximately 45° rotation around the N3-C25-C26-N5 dihedral of the inhibitor down into the protease. This explains the sharp change in R_{cc} observed for the wildtype (see Figure 6.2). The conformational change induced a very stable hydrophobic association formed by the P3 and P2 subsites of the drug, and the P181 and V182 residues of the protease. Additionally a strong hydrogen bond formed between the oxygen atom O^5 of the P3 subsite and the backbone nitrogen atom of D29, with a donor-acceptor distance of less than 3 Å. This explains the reduced fluctuations exhibited by the wildtype after 6 ns, whilst the fact that the rotation was downwards into the protease confirms the reason for the decrease in the Saq-Asp distance observed in Figure 6.1.

Our results confirm the stabilising effect of protonation on the binding event. The flaps in all systems remain closed and the drug remains bound in the active site with a maximum deviation of 3 Å observed for the wildtype. Furthermore, the lateral motion is less than 1 Å away from the original position while the G48V-related lateral dissociation mechanism (reported in Chapter 5) is not observed in the monoprotonated systems. As the monoprotonated system has been reported to be thermodynamically favourable when saquinavir is bound [242] and, from our study, exhibits stable binding in the closed conformation, the G48V mutation is less likely to be able to take advantage of this state to confer resistance. By contrast, the dianionic state exhibits a consistent lateral dissociation mechanism in the G48V and G48V/L90M systems (as discussed in Chapter 5), which is directly due to the additional flap-coupling induced by the G48V mutation. As protonation occurs upon or after ligand binding [156], at physiological pH, the G48V mutation is likely to confer resistance whilst the drug is bound to a dianionic dyad by inducing the first stages of the lateral dissociation, prior to protonation.

However, for the purposes of determining the equilibrium free energy of binding, it is important to use the thermodynamically favourable state. We therefore subsequently conducted free energy calculations using the monoprotonated HIV-1 proteases studied here.

6.3.2 Time-Series and Convergence Analysis of the Enthalpic and Entropic Components of Drug-Binding

Free energies of binding, calculated from MMPBSA, are only reliable if the averages used to compute them have converged. We therefore determined convergence properties by analysing the time evolution of each of the components of the free energy for all drug-protease systems over the 10 ns trajectory (see Figure 6.3). For all systems except the wildtype, all components exhibited stable fluctuations with no observable drift, resulting in the total enthalpic contribution, ΔG_b^{MMPBSA}, having a standard deviation of $\sigma_{sd} < 5$ kcal/mol for each system. Even though the wildtype showed significant variations (~30

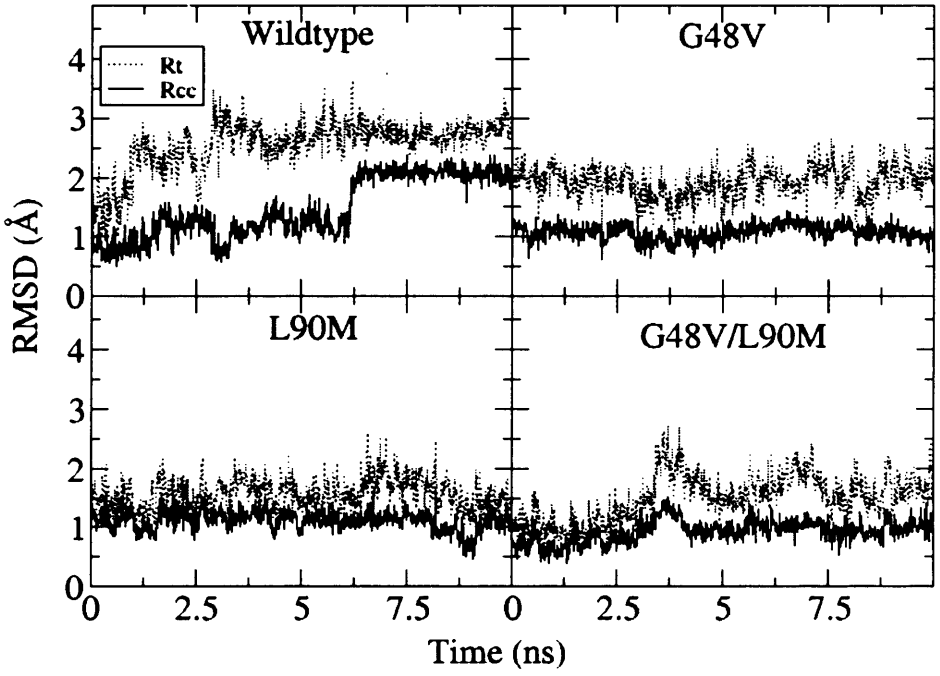


Figure 6.2: Differences in the RMSD of saquinavir relative to its crystal structure in all four monoprotonated protease systems, for two different alignment protocols. R_t (dotted line) shows the RMSD of saquinavir atoms after alignment of the protein backbones, R_{cc} (solid line) after alignment to the heavy atoms of saquinavir. After initial changes in the position of the drug, which in the wildtype lasts for several nanoseconds, all systems exhibit stable inhibitor RMSDs. Unlike the situation pertaining to the dianionic study, there is no observed separation of R_t and R_{cc} here, indicating insignificant 'bulk' motion of the drug.

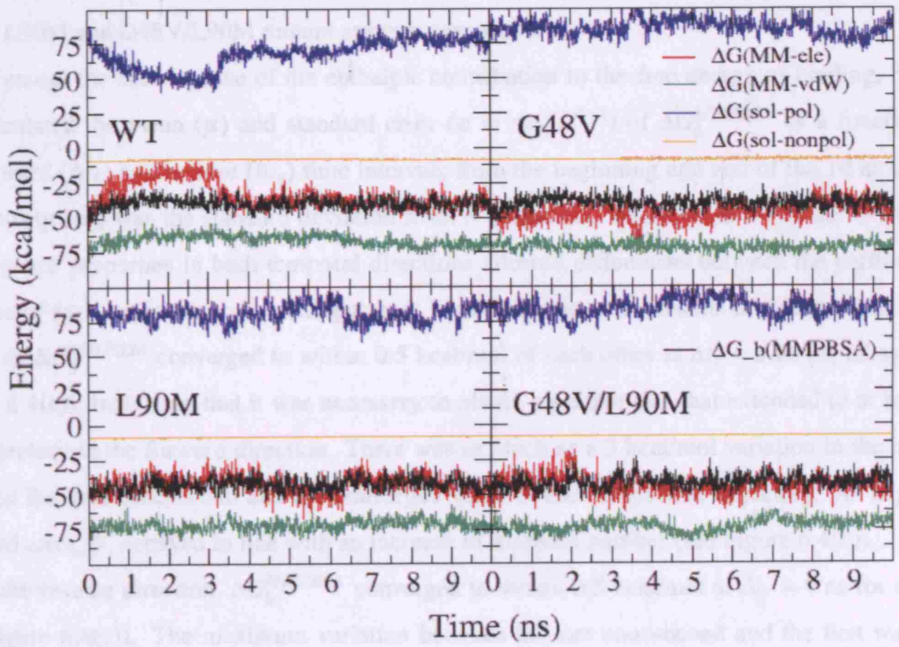


Figure 6.3: Time evolution of each independent component of the MMPBSA contribution to the free energy difference of binding across 1000 snapshots of the 10 ns production run for each drug-protease system. The components of the free energy were stable in all systems, resulting in stable fluctuations of ΔG_b^{MMPBSA} (less than 5 kcal/mol) across the entire 10 ns trajectory. The wildtype exhibited significant variations (~30 kcal/mol) in ΔG_{ele}^{MM} and ΔG_{pol}^{sol} for the first 4 ns. However, anti-correlation of these terms (correlation coefficient: -0.89) resulted in stable fluctuation (~4.5 kcal/mol) of the overall enthalpic contribution.

kcal/mol) in both ΔG_{ele}^{MM} and ΔG_{pol}^{sol} for the first 4 ns of the production run, these two components exhibited significant anti-correlation, such that the total contribution of all components, ΔG_b^{MMPBSA}, still exhibited stable fluctuations across the entire 10 ns trajectory ($\sigma_{sd} \sim 4.5$ kcal/mol). Anti-correlation between this pair of components is expected due to any electrostatic interactions between ligand and protein being compensated by the electrostatic penalty incurred upon solvation of the complex relative to the ligand and receptor separately. The correlation of this pair of terms was therefore determined in each drug-protease system by the regression of the data set across the 10 ns trajectory. This yielded correlation coefficients of -0.89, -0.74, -0.74 and -0.71 between the two components, for the wildtype, G48V, L90M and G48V/L90M mutant systems respectively.

To assess the convergence of the enthalpic contribution to the free energy of binding, ΔG_b^{MMPBSA}, we calculated the mean (μ) and standard error ($\sigma = \sigma_{sd}/N^{1/2}$) of ΔG_b^{MMPBSA} as a function of both the forward (Δt_f) and reverse (Δt_r) time intervals from the beginning and end of the 10 ns trajectories, respectively. σ_{sd} was the standard deviation from N snapshots, where $N/\Delta t = 100$ ns^{-1}. Determining convergence properties in both temporal directions allowed differences between the earlier and latter portions of each trajectory to be distinguished. Figure 6.4 shows the results of this analysis. The mean values of ΔG_b^{MMPBSA} converged to within 0.5 kcal/mol of each other at $\Delta t_f = 5$ ns for all systems (see Figure 6.4(a)), indicating that it was necessary to obtain a sample size that extended to at least 5 ns of the trajectory in the forward direction. There was as much as a 3 kcal/mol variation in the mean value between the first nanosecond and the converged value of each respective trajectory. As expected, the standard error, σ, decayed in line with an increase in snapshot number (see Figure 6.4(b)).

In the reverse direction, ΔG_b^{MMPBSA} converged to within 0.5 kcal/mol at $\Delta t_r = 4$ ns for all systems (see Figure 6.4(c)). The maximum variation between the last nanosecond and the first was approximately 1 kcal/mol. Interestingly, inclusion of the second nanosecond of the wildtype trajectory caused a deviation of $\sim$1 kcal/mol from the converged value, indicating that this portion of the trajectory was not optimal for the assimilation of binding data, although inclusion of the first nanosecond led to convergence below a 0.5 kcal/mol threshold. Again, σ, decayed in line with an increase in snapshot number (see Figure 6.4(d)).

In order to validate the choice of a time interval of 10 ps between successive snapshots, we assessed the autocorrelation for each energy component of a single trajectory over the first 1 ns of production MD, taking 1 snapshot every 1 ps. The G48V HIV-1 protease bound to saquinavir was selected for this analysis. Figure 6.5 shows the autocorrelation of the data set up to a time interval of (a) 30 ps and (b) 500 ps. For a relaxation time based on the drop in correlation by a decay factor of $1/e$, the largest relaxation time occurred between 3-4 ps for ΔG_b^{MMPBSA}, giving a lower bound of 4 ps to the time interval before successive snapshots. The choice of a 10 ps time interval over the 10 ns trajectory therefore yielded both uncorrelated energy values as well as providing a large enough sample size to reduce the standard error of the calculation to below 0.2 kcal/mol in all systems.

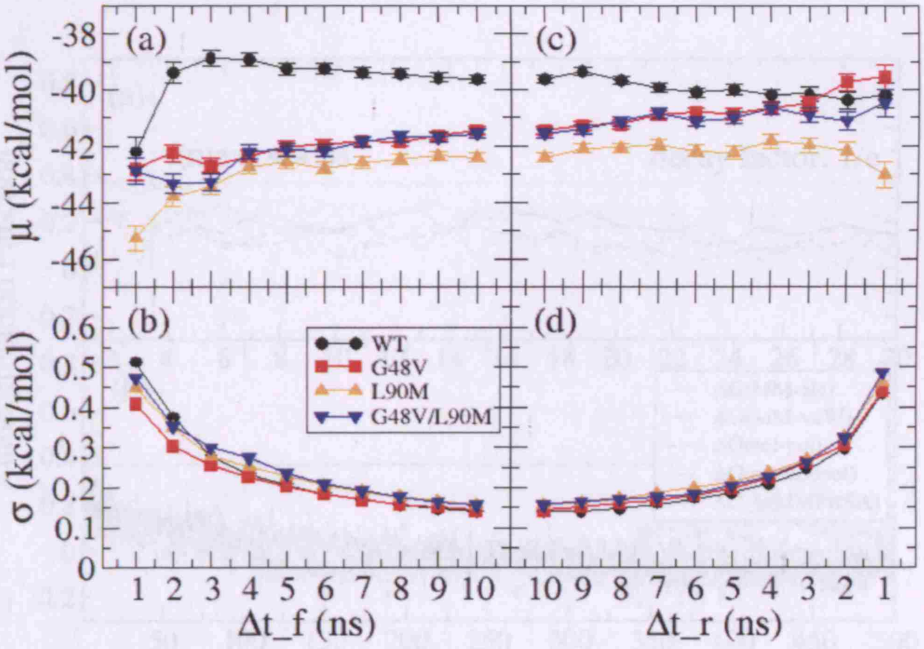


Figure 6.4: Convergence of the enthalpic component of binding, ΔG_b^{MMPBSA}, assessed by (a) the mean (μ) and (b) the standard error ($\sigma = \sigma_{sd}/N^{1/2}$) of the forward ($\Delta t_f$) and (c), (d) reverse ($\Delta t_r$) time intervals, across the 10 ns trajectories for each drug-protease system. σ_{sd} is the standard deviation and N the number of snapshots, where $N/\Delta t = 100$ ns^{-1}. The mean value of ΔG_b^{MMPBSA} converged for all systems within 0.5 kcal/mol at $\Delta t_f = 5$ ns, whilst for the temporally reversed data set, convergence to below a similar threshold occurred at $\Delta t_r = 4$ ns. The standard error for all systems decayed, at the expected $1/N^{1/2}$ rate, with increased N.

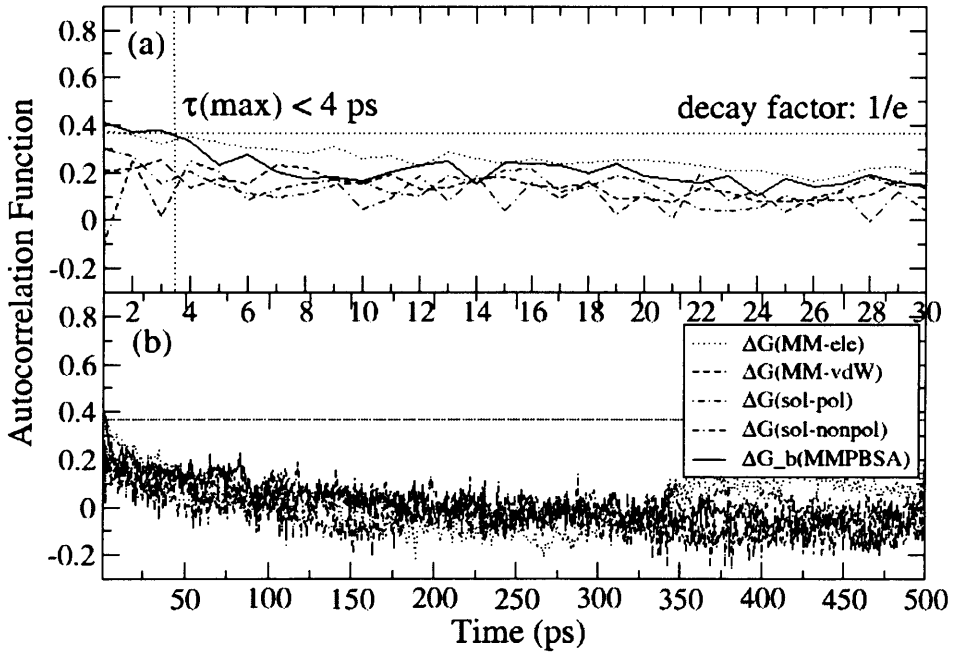


Figure 6.5: Autocorrelation of each independent component of the MMPBSA contribution and the total MMPBSA contribution to the free energy difference of binding across the first 1 ns production run of the G48V system, out to a time interval of (a) 30 ps (b) 500 ps. The largest relaxation time (τ) was between 3-4 ps giving a lower bound to the time interval between successive snapshots of 4 ps. Longer timescale behaviour remained uncorrelated up to the 500 ps interval analysed.

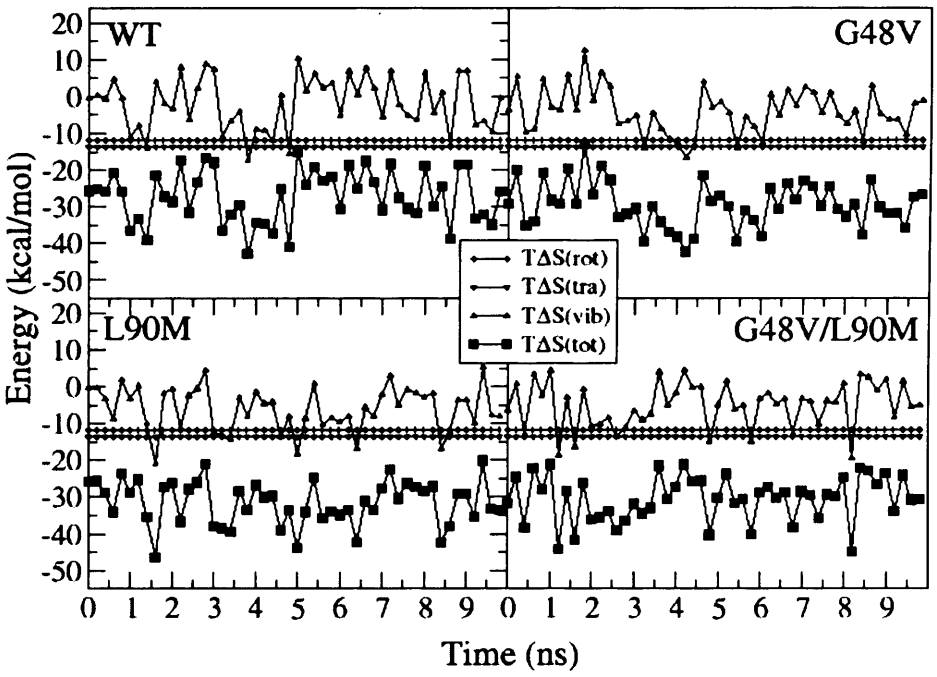


Figure 6.6: Time evolution of the components of configurational entropy, $T\Delta S_{rot}$, for all drug-protease systems. 50 equally spaced snapshots were selected across the 10 ns trajectory. The values of $T\Delta S_{rot}$ and $T\Delta S_{tra}$ were effectively constant across all trajectories. Large variation with a range of $\sim$ 25 kcal/mol was observed for $T\Delta S_{vib}$ as well as fluctuations with a standard deviation of approximately 7 kcal/mol.

The variation of the components of the configurational entropy contribution for all systems, assessed by normal mode analysis, are shown in Figure 6.6. Whilst the translational ($T\Delta S_{tra}$) and rotational ($T\Delta S_{rot}$) components of the entropy were well-behaved and effectively constant, there was substantial variation in the vibrational ($T\Delta S_{vib}$) component of the entropy across the different snapshots, with a range of $\sim$ 25 kcal/mol and a standard deviation of approximately 7 kcal/mol.

The convergence of the entropic component of binding, $T\Delta S$, was determined in the same way to that described for the enthalpy (see Figure 6.7) with the exception that the number of snapshots used was much smaller across the 10 ns trajectory ($N/\Delta t$ = 5 ns^{-1}). The mean values of $T\Delta S$ converged to within 0.5 kcal/mol of each other at Δt_f = 6 ns for all systems (see Figure 6.7(a)). There was as much as a 3 kcal/mol variation in the mean value between the first nanosecond and the converged value of each respective trajectory. The standard error, σ, for the first 4 ns showed deviation from an expected decay rate (see Figure 6.7(b)), followed by reversion to normal decay with an increase in snapshot number.

In the temporally reversed direction, $T\Delta S$ also converged to within 0.5 kcal/mol at Δt_r = 6 ns for all systems (see Figure 6.7(c)), whilst the maximum variation between the last nanosecond and the first was approximately 2 kcal/mol. Incidentally, the value for the entropy after 4 ns of sampling was within the threshold for convergence. However, due to an $\sim$1 kcal/mol deviation exhibited after 5 ns, the data set did not converge until 6 ns into production. Similar deviations were observed for the decay of σ in the initial 3 ns, followed by normal decay with an increase in N (see Figure 6.7(d)).

Overall, our analysis showed that sampling of at least 4 ns of the 10 ns trajectory, at a rate of $N/\Delta t$ = 100 ns^{-1}, was necessary to obtain converged enthalpies of binding and that at least 6 ns of sampling, with $N/\Delta t$ = 5 ns^{-1}, were necessary for the convergence of the entropy. Convergence analysis showed that for both the enthalpy and the entropy, averaged-energies derived from the latter portion of each trajectory were closer to the converged values than those derived from the earlier portions. This was likely due to further structural readjustments well into the post-equilibration phase, (see § 6.3.1), and supported the notion that the entire 10 ns trajectory was not preferable to the latter several nanoseconds of each trajectory in the calculation of the free energy of binding.

6.3.3 Absolute and Relative Free Energy Differences of Binding of Saquinavir to HIV-1 Proteases

We investigated the efficacy of our approach in obtaining accurate absolute free energy differences of binding as well as the sensitivity of our approach in distinguishing the binding affinity of saquinavir to the wildtype protease and the G48V, L90M and G48V/L90M mutants. The entire 10 ns of each trajectory was selected for the analysis (see Table 6.1). However, as both structural and convergence analyses indicated conformational readjustments in the drug-wildtype protease complex, which did not stabilise until just under 6 ns post-equilibration, we also report values from the last 4 ns of each trajectory

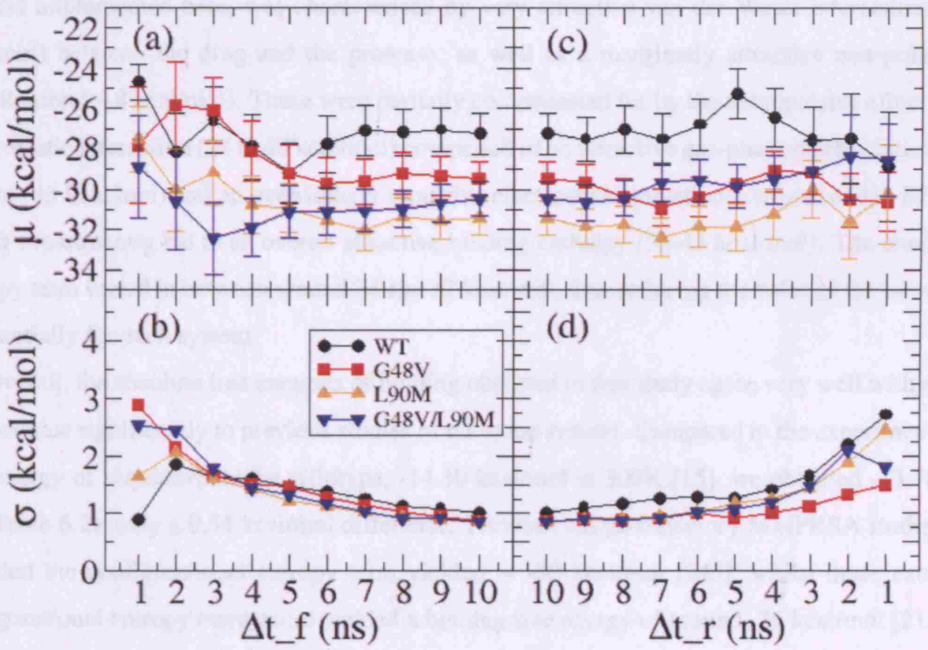
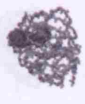


Figure 6.7: Convergence of the entropic component of binding $T\Delta S$, assessed by (a) the mean (μ) and (b) the standard error ($\sigma = \sigma_{sd}/N^{1/2}$) of the forward ($\Delta t_f$) and (c),(d) reverse ($\Delta t_r$) time intervals across the 10 ns trajectories for each drug-protease system. σ_{sd} is the standard deviation and N the number of snapshots, where $N/\Delta t = 5$ ns^{-1}. The mean value of $T\Delta S$ converged for all systems within 0.5 kcal/mol at $\Delta t_f = 6$ ns and similarly convergence of the temporally reversed data set to below a similar threshold occurred at $\Delta t_r = 6$ ns. For the first 4 ns and 5 ns in the forward and reverse directions respectively, the standard error deviated from the expected decay rate of $1/N^{1/2}$, followed by reversion to an expected decay with increased N.



(see Table 6.2). Furthermore, to determine the extent to which shorter trajectories reproduced the relative and absolute binding profiles exhibited by larger converged trajectories, we included the results from only the first 1 ns (see Table 6.3). In order to facilitate a better comparison between the three systems, the entropy from this last data set was calculated separately from 20 equally spaced snapshots across the 1 ns trajectory. We compared our results against two experimental sets of data for the same mutants, denoted ε_1 [15] and ε_2 [194].

The binding of saquinavir to the wildtype and all mutant proteases, as calculated by the MMPBSA method implemented here, was characterised by very attractive van der Waals interactions (66 to 74 kcal/mol) between the drug and the protease, as well as a marginally attractive non-polar solvation contribution ($\sim$ 8 kcal/mol). These were partially compensated for by the net repulsive effect of the total electrostatic interaction (34 to 40 kcal/mol) comprised of an attractive gas-phase electrostatic component (-30 to -45 kcal/mol) and an overridingly repulsive electrostatic solvation component (65-87 kcal/mol), which subsequently led to an overall attractive binding enthalpy (39-45 kcal/mol). The configurational entropy term varied in between around 24 and 32 kcal/mol, thus reducing the value of the overall binding substantially for each system.

Overall, the absolute free energies of binding obtained in this study agree very well with experiment and contrast significantly to previous studies of the same system. Compared to the experimental binding free energy of saquinavir to the wildtype, -14.30 kcal/mol at 300K [15], we obtained -13.76 kcal/mol (see Table 6.2), only a 0.54 kcal/mol difference. Previous single-trajectory MMPBSA studies that also included the configurational entropy term yielded $\sim$ -30 kcal/mol [245], whilst those excluding the configurational entropy component yielded a binding free energy of around -26 kcal/mol [218].

Our study supports the notion that the loss of configurational entropy upon binding plays a significant role in the overall absolute free energy of binding. Recent studies have shown that the configurational entropy contribution of ligand-binding to HIV-1 protease is non-negligible and has a value of around 25 kcal/mol [267] for amprenavir, which agrees well with the range of 24 to 32 kcal/mol obtained in our studies.

The three sets of time-averaged values differed only slightly from each other. The largest differences between calculated and experimental binding free energies were $\sim$3 kcal/mol, $\sim$2 kcal/mol and $\sim$1 kcal/mol, for the 1 ns, 10 ns and 4 ns time-averaged data sets respectively (see Tables 6.1, 6.2 and 6.3). The closest agreement with experiment ε_1, obtained from the 4ns time-averaged analysis of the G48V mutant, was within 0.03 kcal/mol from the experimental value. Furthermore, the correct order of the relative binding strengths of the different HIV-1 protease variants was only obtained in the 4 ns trajectory. It is likely, therefore, that even though the 10 ns sample was larger and had converged, conformational readjustments in the systems, particularly the wildtype, in the earlier stages of the production run, impeded more accurate results from being determined. By contrast the 4 ns averages avoided the earlier stages of the trajectory, but were still sufficiently large to have converged. The 1 ns averages

Enthalpic components of binding ($\Delta t_{f/r}$ = 10 ns)

Sequence	ΔG^{MM}_{vdW}	ΔG^{MM}_{ele}	ΔG^{sol}_{nonpol}	ΔG^{sol}_{pol}	ΔG^{tot}_{ele}	ΔG^{MMPBSA}_{b}
WT	**-66.12** (0.141)	**-33.24** (0.307)	**-7.98** (0.008)	**67.71** (0.330)	**34.48** (0.150)	**-39.63** (0.140)
G48V	**-71.59** (0.108)	**-47.34** (0.218)	**-8.42** (0.005)	**85.91** (0.191)	**38.57** (0.150)	**-41.44** (0.142)
L90M	**-71.63** (0.129)	**-42.50** (0.222)	**-8.46** (0.005)	**80.18** (0.203)	**37.68** (0.153)	**-42.41** (0.157)
G48V/L90M	**-72.93** (0.145)	**-43.94** (0.227)	**-8.36** (0.005)	**83.66** (0.212)	**39.72** (0.168)	**-41.57** (0.157)

Entropic components and absolute free energies of binding ($\Delta t_{f/r}$ = 10 ns)

Sequence	$T\Delta S_{tra}$	$T\Delta S_{rot}$	$T\Delta S_{vib}$	$T\Delta S_{tot}$	ΔG_b	$\Delta G(\varepsilon_1)$	$\Delta G(\varepsilon_2)$
WT	**-13.58** (0.000)	**-11.80** (0.003)	**-1.85** (1.015)	**-27.24** (1.017)	**-12.39** (1.157)	**-14.30** (0.084)	**-13.76** (0.162)
G48V	**-13.58** (0.000)	**-11.83** (0.003)	**-4.10** (0.878)	**-29.51** (0.878)	**-11.93** (1.020)	**-11.28** (0.062)	**-12.16** (0.044)
L90M	**-13.58** (0.000)	**-11.83** (0.001)	**-6.06** (0.851)	**-31.47** (0.853)	**-10.94** (1.010)	**-12.51** (0.034)	**-13.09** (0.040)
G48V/L90M	**-13.58** (0.000)	**-11.82** (0.003)	**-5.10** (0.862)	**-30.51** (0.862)	**-11.06** (1.019)	**-10.21** (0.018)	**-10.04** (0.140)

Mean energies are in kcal/mol, corresponding standard errors in parentheses.

Enthalpic sample size: $N/\Delta t$ = 100 ns^{-1}, entropic sample size: $N/\Delta t$ = 5 ns^{-1}

$\Delta G(\varepsilon_1)$ [15] and $\Delta G(\varepsilon_2)$ [194] are experimental results converted from inhibition constants at T = 298.15 K and T = 310.15 K respectively.

Table 6.1: Enthalpic and entropic decomposition of saquinavir binding to HIV-1 protease wildtype and mutants time-averaged over all 10 ns.

Enthalpic components of binding (Δt_r = 4 ns)

Sequence	ΔG^{MM}_{vdW}	ΔG^{MM}_{ele}	ΔG^{sol}_{nonpol}	ΔG^{sol}_{pol}	ΔG^{tot}_{ele}	ΔG^{MMPBSA}_{b}
WT	**-68.89** (0.181)	**-39.29** (0.228)	**-8.08** (0.010)	**76.05** (0.242)	**36.76** (0.225)	**-40.20** (0.211)
G48V	**-72.12** (0.166)	**-43.95** (0.336)	**-8.38** (0.009)	**83.78** (0.296)	**39.83** (0.231)	**-40.66** (0.220)
L90M	**-70.53** (0.210)	**-42.03** (0.337)	**-8.48** (0.007)	**78.20** (0.334)	**36.17** (0.220)	**-42.84** (0.237)
G48V/L90M	**-70.66** (0.192)	**-44.55** (0.281)	**-8.31** (0.009)	**82.86** (0.290)	**38.32** (0.229)	**-40.66** (0.227)

Entropic components and absolute free energies of binding (Δt_r = 4 ns)

Sequence	$T\Delta S_{tra}$	$T\Delta S_{rot}$	$T\Delta S_{vib}$	$T\Delta S_{tot}$	ΔG_b	$\Delta G(\varepsilon_1)$	$\Delta G(\varepsilon_2)$
WT	**-13.58** (0.000)	**-11.81** (0.004)	**-1.05** (1.447)	**-26.44** (1.449)	**-13.76** (1.660)	**-14.30** (0.084)	**-13.76** (0.162)
G48V	**-13.58** (0.000)	**-11.83** (0.004)	**-3.94** (1.042)	**-29.35** (1.044)	**-11.31** (1.264)	**-11.28** (0.062)	**-12.16** (0.044)
L90M	**-13.58** (0.000)	**-11.83** (0.002)	**-5.88** (1.288)	**-31.29** (1.290)	**-11.55** (1.527)	**-12.51** (0.034)	**-13.09** (0.040)
G48V/L90M	**-13.58** (0.000)	**-11.83** (0.004)	**-4.61** (1.230)	**-30.02** (1.232)	**-10.64** (1.459)	**-10.21** (0.018)	**-10.04** (0.140)

Mean energies are in kcal/mol, corresponding standard errors in parentheses.

Enthalpic sample size: $N/\Delta t$ = 100 ns^{-1}, entropic sample size: $N/\Delta t$ = 5 ns^{-1}.

$\Delta G(\varepsilon_1)$ [15] and $\Delta G(\varepsilon_2)$ [194] are experimental results converted from inhibition constants at T = 298.15 K and T = 310.15 K respectively.

Table 6.2: Enthalpic and entropic decomposition of saquinavir binding to HIV-1 protease wildtype and mutants time-averaged over the last 4 ns.

Enthalpic components of binding (Δt_f = 1 ns)

Sequence	ΔG^{MM}_{vdW}	ΔG^{MM}_{ele}	ΔG^{sol}_{nonpol}	ΔG^{sol}_{pol}	ΔG^{tot}_{ele}	ΔG^{MMPBSA}_{b}
WT	**-68.50** (0.439)	**-32.90** (0.916)	**-8.35** (0.015)	**67.53** (0.753)	**34.63** (0.479)	**-42.22** (0.515)
G48V	**-70.56** (0.336)	**-50.02** (0.573)	**-8.45** (0.018)	**86.20** (0.458)	**36.18** (0.413)	**-42.82** (0.410)
L90M	**-71.93** (0.390)	**-44.52** (0.597)	**-8.43** (0.019)	**79.57** (0.490)	**35.05** (0.472)	**-45.30** (0.451)
G48V/L90M	**-73.72** (0.369)	**-42.67** (0.689)	**-8.41** (0.014)	**81.90** (0.501)	**39.22** (0.452)	**-42.91** (0.487)

Entropic components and absolute free energies of binding (Δt_f = 1 ns)

Sequence	$T\Delta S_{tra}$	$T\Delta S_{rot}$	$T\Delta S_{vib}$	$T\Delta S_{tot}$	ΔG_b	$\Delta G(\varepsilon_1)$	$\Delta G(\varepsilon_2)$
WT	**-13.58** (0.000)	**-11.81** (0.002)	**0.70** (1.190)	**-24.69** (1.190)	**-17.53** (1.705)	**-14.30** (0.084)	**-13.76** (0.162)
G48V	**-13.58** (0.000)	**-11.82** (0.002)	**-3.90** (1.476)	**-29.30** (1.476)	**-13.52** (1.886)	**-11.28** (0.062)	**-12.16** (0.044)
L90M	**-13.58** (0.000)	**-11.83** (0.004)	**-3.79** (1.176)	**-29.20** (1.176)	**-16.10** (1.627)	**-12.51** (0.034)	**-13.09** (0.040)
G48V/L90M	**-13.58** (0.000)	**-11.81** (0.002)	**-3.31** (1.411)	**-28.70** (1.411)	**-14.21** (1.898)	**-10.21** (0.018)	**-10.04** (0.140)

Mean energies are in kcal/mol, corresponding standard errors in parentheses.

Enthalpic sample size: $N/\Delta t = 100\ \text{ns}^{-1}$, entropic sample size: $N/\Delta t = 20\ \text{ns}^{-1}$.

$\Delta G(\varepsilon_1)$ [15] and $\Delta G(\varepsilon_2)$ [194] are experimental results converted from inhibition constants at T = 298.15 K and T = 310.15 K respectively.

Table 6.3: Enthalpic and entropic decomposition of saquinavir binding to HIV-1 protease wildtype and mutants time-averaged over the first 1 ns.

were the least accurate; however it is nonetheless surprising that such a high level of agreement with experiment was achievable with such a modest sample size.

In order to investigate the suitability of our method in effectively ranking drug resistant mutants of HIV-1 protease, we investigated the relative free energies of binding of the wildtype and mutant proteases against the two experimental sets of data for the same protease variants. Figure 6.8 shows the correlation between the experimental values and the ranking obtained from using just the MMPBSA energies as well as from the inclusion of configurational entropies, for all three time-averaged data sets.

Using MMPBSA alone (see Figure 6.8(a)), no significant correlation was observed between any of the computed data sets and either of the experimental values. Indeed, the 10 ns trajectory showed signs of anti-correlation with correlation coefficients of -0.62 and -0.39 to data sets ε_1 and ε_2 respectively. This was primarily due to the L90M mutant, which showed a more attractive enthalpy by 2-3 kcal/mol than the wildtype across all three computed data sets. Furthermore, the G48V and G48V/L90M mutants also exhibited marginally more attractive enthalpies (0.4-2 kcal/mol) than the wildtype.

Upon inclusion of the configurational entropy, the relative ranking improved substantially. The best ranking was obtained for the 4 ns trajectory, with correlation coefficients of 0.96 and 0.81 with respect to experimental data sets ε_1 and ε_2 respectively. These are remarkably good, when considering that the correlation coefficient between the two experimental data sets themselves was 0.93. Furthermore, the binding of saquinavir to each mutant protease was correctly ordered relative to the wildtype. It should be noted that a main factor in the discrepancy between the experimental results ε_1 and ε_2 is the difference in temperature at which the experiments were conducted (298.15 K and 310.15 K respectively). In the study reported here, the mean equilibrium temperature was 299.23 K for all systems, making our results more directly comparable with ε_1 than ε_2. It is therefore additionally encouraging that a greater correlation coefficient was exhibited relative to ε_1 than ε_2.

Unsurprisingly, rescaling of the binding free energies attained for the 4 ns trajectory to 310.15 K resulted in a deterioration in the correlation coefficients, specifically to 0.88 and 0.69 for ε_1 and ε_2 respectively. This is because only the configurational entropy component ($T\Delta S_{tot}$) can be rescaled easily, whilst the solvation entropy component is incorporated in the MMPBSA calculation performed at approximately 299 K. An accurate comparison with ε_2 would thus require simulation at the higher temperature of 310.15 K.

Interestingly, better ranking was achieved using the 1 ns trajectories than the long 10 ns trajectories, with correlation coefficients of $Cc(\varepsilon_1) = 0.91$, $Cc(\varepsilon_2) = 0.75$ and $Cc(\varepsilon_1) = 0.60$, $Cc(\varepsilon_2) = 0.53$ for the two data sets respectively. However, based on the convergence analysis conducted in § 6.3.2, in which the 1 ns trajectory was shown not to have converged, it is likely that this is something more of a fortuitous result. Additionally the inclusion of the earlier portions of the trajectory in which conformational changes of the drug were still occurring, especially in the wildtype system, are likely to have adversely affected the statistical time-averaging over the 10 ns trajectory. As the 4 ns time-average was based on

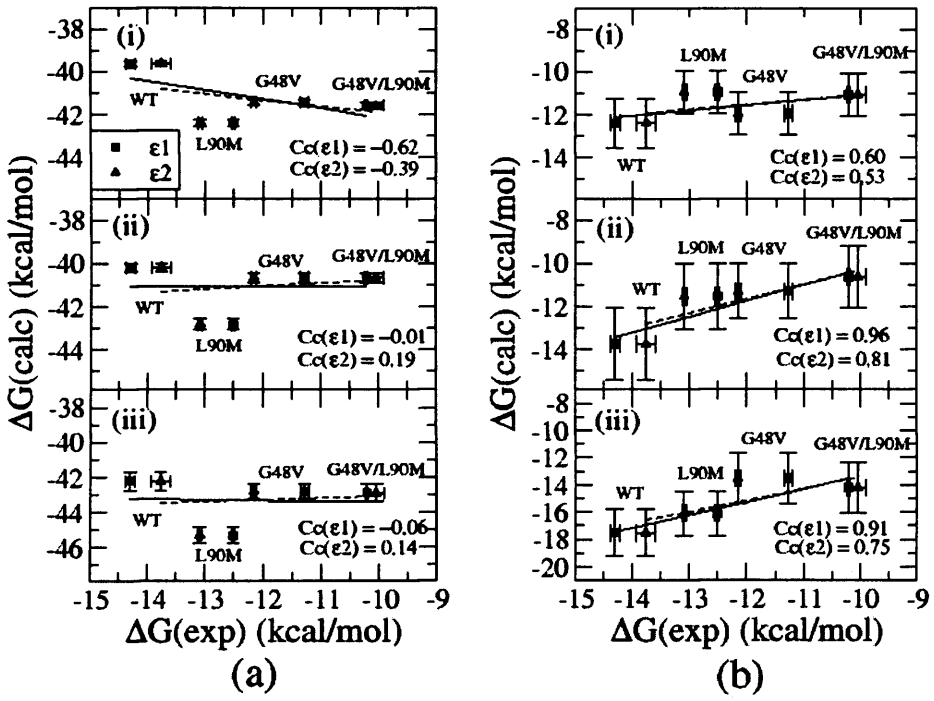


Figure 6.8: Correlation of relative free energies of binding to experiment for wildtype HIV-1 protease and the G48V, L90M and G48V/L90M mutants. Correlations were evaluated using (a) MMPBSA alone, (b) including configurational entropy for the (i) 10 ns, (ii) 4 ns and (iii) 1 ns time-averaged data sets. Calculated values were compared against two experimental data sets ε_1 and ε_2 from [15] and [194] respectively. Binding free energies were converted from inhibition constants at T = 298.15 K and T = 310.15 K for ε_1 and ε_2 respectively. The best ranking is achieved for the 4 ns trajectory upon inclusion of configurational entropy.

the last 4 ns, such conformational changes were avoided, resulting in a remarkably close correlation with experiment.

The 4 ns trajectory was then used for subsequent analysis of the basis of resistance in each of the mutants (see Table 6.2). The effect of G48V both in the single and double mutant was to cause a marginal increase in the binding enthalpy (ΔG_b^{MMPBSA} ~0.5 kcal/mol). This was, as expected, due to the valine mutation interacting with the P3 subsite of saquinavir, leading to an increased van der Waals interaction (~2-3 kcal/mol) in the binding site, which was only slightly compensated by an increased solvation penalty (~1.5-2.5 kcal/mol). However the L90M mutant exhibited a 2.5 kcal/mol more attractive enthalpy than the wildtype. This was comprised of an ~2 kcal/mol more attractive van der Waals interaction, which was not compensated by an increased solvation penalty as the L90M mutation is buried in the protease and not in the active site. Instead, the solvation penalty decreased by ~0.5 kcal/mol.

All mutants substantially affected the entropic penalty of binding ($T\Delta S_{tot}$), with increases of approximately 3, 5 and 4 kcal/mol for the G48V, L90M and G48V/L90M mutants respectively. The net effect was to reduce the overall binding of the G48V and G48V/L90M mutants by over 2 and 3 kcal/mol relative to the wildtype respectively. The substantially increased entropic penalty in the L90M mutant served to attenuate the increased enthalpy of binding causing a ~1 kcal/mol reduction in overall binding. Therefore, whilst G48V resistance is driven by a slightly increased entropic barrier alone, L90M resistance is driven by larger entropic penalties to binding, which are moderated by slight increases in binding enthalpy.

Care must be taken however, when interpreting these results. The standard error in the overall determination of the free energy difference was approximately 1.5 kcal/mol, owing largely to the uncertainty in the entropy, even with a relatively large sample size of N=50. Therefore, an alteration in the relative ranking may occur upon further increasing the sample size. Nevertheless, as the values of the entropy were not significantly affected in the transition from the 4 ns to the 10 ns trajectory analysis, it is likely that more sampling would reduce the standard error without any further significant change in the binding characteristics.

To summarise, our findings reveal that calculation of the change in configurational entropy upon binding, alongside an assessment in the enthalpic change from MMPBSA calculations, is essential for the determination of accurate binding free energies. We have pursued such a strategy to successfully determine accurate free energies of binding of saquinavir to the wildtype and G48V, L90M and G48V/L90M mutant HIV-1 proteases. Furthermore, inclusion of configurational entropies substantially improves the relative ranking profile of drug resistant mutants of HIV-1 protease relative to the wildtype and allows mutants with as little as 1.5 kcal/mol deviation from the wildtype to be distinguished. However, owing to the sensitivity of the entropy as well as the enthalpy calculations, trajectories must be suitably long (ideally more than 4 ns) and sampled frequently enough to ensure convergence of the

time-averaged values of the free energy, as well as significant reduction of the standard error. Finally, sampling should only begin once structural readjustments have stabilised.

6.4 Conclusion

Calculating absolute binding free energies is still a computational challenge in biological applications [245]. The MMPBSA approximate method has been applied in the past to several HIV-1 protease inhibitor complexes [218, 245, 260]. However, considerable effort is still needed in validating the effectiveness of the method for discriminating among drug-resistant protein mutations and for elucidating the molecular basis of drug resistance on clinically relevant time scales.

In the present study we report fully unrestrained molecular dynamics, combined with MMPBSA thermodynamic analysis of saquinavir binding to the wildtype and three saquinavir-resistant protease mutants, G48V, L90M and G48V/L90M. We calculated relative and absolute binding affinities for saquinavir and obtained an excellent level of correlation with experimental values [15, 194]. Absolute values of binding were determined within 1 kcal/mol of the experimental results and the relative binding free energies of the mutants considered here were correctly ordered with respect to the wildtype. Our study thus achieves a significant advance over previous studies that have implemented the MMPBSA method on drug-bound HIV-1 proteases [245] and is the first study to successfully rank protease mutants using this method.

Our findings reveal that the inclusion of configurational entropy is essential to provide an accurate value for the absolute binding affinity of ligand-bound HIV-1 proteases and the values of configurational entropy, determined here, are in good agreement with previous studies [267]. We have additionally determined convergence properties of both enthalpic and entropic time-averages across a 10 ns timescale. Our study indicates that several nanoseconds of MD with frequent sampling of snapshots are required for suitable convergence. Furthermore, even after a considerable period of time designated for equilibration (2 ns in this study), the actual equilibration of systems may extend well into the earlier portions of trajectories intended for the production phase. Ideally, sampling should thus begin once such structural readjustments have stabilised.

Additionally, our study shows that inclusion of configurational entropy substantially improves the relative ranking of drug resistant mutants, allowing for a clear differentiation of mutants with binding free energy differences upwards of 1.5 kcal/mol. Indeed, our findings show that the overall resistance exhibited by a mutant cannot necessarily be explained by just enthalpic penalties; instead the interplay of enthalpic and entropic gains and losses is required to correctly describe the basis of resistance.

Based on the enthalpic and entropic decomposition of our analysis, resistance induced by the G48V mutation is driven by a slightly increased entropic barrier to binding, whilst enthalpy is largely similar to the wildtype. L90M resistance is driven by a substantially increased entropic barrier compensated for

by a slightly more favourable enthalpic contribution to binding. In the G48V/L90M double mutant, the G48V mutation keeps the enthalpy similar to that of the wildtype, whilst the presence of L90M, just as is the case in the single L90M mutant, incurs a similarly increased entropic penalty.

Finally we have developed a high-throughput tool, called the 'Binding Affinity Calculator' (BAC) for the automated calculation of binding free energies of both inhibitors and natural substrates binding to wildtype and mutant HIV-1 proteases. The study presented here of saquinavir and its characteristic drug-resistant mutants, as well as the study presented in Chapter 7 of the NC-p1 substrate binding to wildtype and mutant HIV-1 proteases, serve as examples of how the BAC has been used to handle workflows involved in the calculation of binding free energies. Furthermore, the BAC can be integrated into decision support systems used in the clinical environment, to complement information regarding the drug resistance conferred by specific mutant strains of the protease.

CHAPTER 7

Towards a Ranking of the Enzymatic Fitness of HIV-1 Proteases using Free Energy Methods

THE fitness of a particular strain of a virus is a measure of its ability to survive in its biological environment. The biological environment is always subject to change and retroviruses, as discussed in Chapter 3, take advantage of their high mutational rate to evolve and thus adapt to such changes. In the context of HIV, the application of anti-retroviral inhibitors, designed to inhibit the wildtype enzyme, induces a change in the environment of the virus and natural selection allows drug-resistant mutants to proliferate. The subsequent loss of fitness of the wildtype virus can thus be partially restored by evolution to a mutant strain that is fitter in this chemotherapeutic environment than the wildtype.

As most inhibitors have specific enzymatic targets, for example the HIV-1 protease, it is not surprising that mutations in such enzymes are correlated with their respective inhibitors (see Chapter 3). It then becomes possible to discuss the effect of mutations on the corresponding enzymatic fitness, the overall rate of enzymatic function in the presence of inhibitors. It is generally assumed that persistent mutations in HIV-1 protease, in response to drug treatment, increase the enzymatic fitness of the protease. Furthermore, it is thought that the pathway of acquired mutations correlates with a steady increase in enzymatic fitness in response to the chemotherapeutic environment (see Chapter 3).

In this chapter, our aim is to investigate the effects of mutations on both the catalytic efficiency and the enzymatic fitness of HIV-1 protease using free energy methods based on molecular simulation.

7.1 Background

Ultimately, it is the overall viral fitness of a particular sequence that directs its persistence *in vivo* [222]. The viral fitness is in turn dependent on the phenotypic fitness of the array of enzymes and proteins utilised for its propagation. Chemotherapeutic pressure reduces the enzymatic fitness of the wildtype protease due to competition between the inhibitor and naturally cleaved substrates. The subsequent change in such enzymatic fitness due to a mutation in, for example, the HIV-1 protease is not only a

function of how well the protease resists drug binding, but also of the change in catalytic efficiency induced by the mutation. As discussed in Chapter 3, it is this interplay between the gains and losses of both catalytic efficiency and drug resistance which determines whether a mutation will be selected [204, 224]. Previous studies have shown the deleterious effects of some drug-resistant mutations on the catalytic efficiency of the protease [227]. Nevertheless their emergence in response to chemotherapeutic pressure supports an overriding reduction in drug binding as compared to a reduction in binding of the natural substrates [223]. Compensatory mutations have also been shown to increase enzymatic fitness both in the presence and absence of inhibitor pressure [228].

As discussed in Chapter 3, the protease cleaves the Gag and Gag-Pol polyprotein chains at ten distinctly recognised cleavage sites [129]. Such broad specificity is modulated by differential rates of processing for these sites, the slowest rates being those of the NC-p1 (amino acid sequence: RQAN-FLGK) and the Ca-p2 (amino acid sequence: ARVL-AEAM) substrates [229]. The effect of mutations on altering the catalytic efficiency with which these substrates are processed is likely to be a significant parameter in the overall change in the viral fitness of a mutant. However, the overriding importance of understanding changes in the processing efficiency of NC-p1 is supported by the fact that Gag polyprotein mutations, which compensate losses in catalytic efficiency incurred by primary drug resistant mutations, have been reported for this substrate [229, 230, 237]. It then follows that protease mutations that enhance the catalytic efficiency of NC-p1 processing may be selected, provided they result in increased enzymatic fitness in the presence of the inhibitor. Indeed, the L90M mutation has been found to increase the catalytic efficiency for a range of substrates [15], including NC-p1 [230].

Whilst computational studies, conducted by others and discussed in Chapter 6, have been invaluable in explaining some of the molecular characteristics of resistance mutations, they have not often included the effect of mutations on the catalytic efficiency or the subsequent enzymatic fitness of the protease. Indeed, the lack of crystal structures of the protease bound to natural substrates has compounded the difficulty of subsequent computational modelling. Some studies have been performed on the natural substrates cleaved by the protease, exploring both the role of mutations on flexibility as well as the actual catalytic mechanism [133, 139]. These studies involved conversion of peptidomimetic inhibitors into substrates first. Recently however, several structures of the naturally cleaved substrates bound to inactive protease have been crystallised [128, 129, 131], thus facilitating a comparative study between the effect of a mutation on inhibitor binding as compared to substrate binding.

In this study, we formulate a metric for ranking not only the drug resistance conferred by mutants, but also the overall fitness of mutant enzymes whilst under chemotherapeutic pressure. Then we use molecular dynamics simulations in explicit water alongside the MMPBSA method, including entropic considerations from normal mode analysis, to determine the absolute free energies of binding of the NC-p1 cleavage substrate of the Gag polyprotein to the wildtype, the G48V, L90M and G48V/L90M mutants of HIV-1 protease. Furthermore, we explore the applicability of the MMPBSA approach, combined with

analysis of the entropic contribution, in improving the accuracy of absolute free energies of binding as well as investigating the convergence of free energy results for the substrate as compared to the inhibitor saquinavir presented in Chapter 6.

Finally, by combining the effects of a set of mutants on drug binding affinities, determined in Chapter 6, with their corresponding effects on the binding of a natural substrate to HIV-1 protease, we extend our analysis to a ranking of the enzymatic fitness of the mutations under the selection pressure of saquinavir. We are thus able to discuss the applicability of our formulated metric in identifying the pathway of acquired mutations, in response to inhibitor pressure.

7.2 Theoretical Considerations: The Free Energy Potential of Enzymatic Fitness

The enzymatic fitness of a mutant *in vivo* under chemotherapeutic pressure is dependent not only on its effect on inhibitor binding but also on the subsequent changes in catalytic efficiency. An approach that combines the two for a particular set of mutants is more indicative of the overall resistance ranking for those mutants *in vivo*.

Experimentally, the effect of a mutant on inhibitor binding can be determined by the ratio of the inhibition constants of the mutant and wildtype proteases ($K_{i,mut}/K_{i,wt}$). The catalytic efficiency of a particular substrate is determined by the ratio of the kinetic parameters for catalysis (k_{cat}/K_m).

Gulnik *et al.* [224] devised a *vitality* metric (**V**) which effectively described the interplay of the mutant-induced changes in inhibitor binding with changes in catalytic efficiency (see Equation 7.1).

$$\mathbf{V} = \frac{(K_i.k_{cat}/K_m)_{mut}}{(K_i.k_{cat}/K_m)_{wt}} \tag{7.1}$$

Within such a scheme, the higher the vitality, the higher the probability of a mutant to be selected under inhibitor pressure. Such a metric has also been used by others to describe the emergence of several drug resistant mutations [204, 225].

Here we represent this metric in terms of the binding free energy difference between inhibitor and substrate. Interestingly, differences in substrate and inhibitor binding free energies, decomposed by amino acid residue, have been investigated before to predict amino acids susceptible to mutation in the wildtype protease [218]. However, such differences were not applied globally to a set of mutants, nor correlated with the established vitality metric devised by Gulnik *et al.* [224] in describing enzymatic fitness.

Previous studies have shown that k_{cat} does not significantly change with mutation, whilst the vitality can change by several orders of magnitude [15, 224]. The main factor in determining the vitality of a mutant is then the change in the ratio K_i/K_m and we can therefore make the assumption that

$(k_{cat})_{mut}/(k_{cat})_{wt} \sim 1$. Rewriting the binding constants in terms of free energies of inhibitor (ΔG_{drug}) and substrate (ΔG_{sub}) binding we have:

$$\frac{K_i}{K_m} = \exp(-\frac{1}{RT}(\Delta G_{sub} - \Delta G_{drug})) \tag{7.2}$$

where R is the gas constant and T the temperature. We define a new metric $V_f(\mu)$, the *free energy potential of enzymatic fitness* as the difference between substrate and inhibitor binding:

$$V_f(\mu) = \Delta G_{sub}(\mu) - \Delta G_{drug}(\mu) \tag{7.3}$$

where μ varies over the mutational landscape of the protease. The vitality of a mutant (*mut*) with respect to the wildtype (*wt*) can subsequently be expressed in terms of this metric as:

$$\mathbf{V} = \frac{k_{cat}(mut)}{k_{cat}(wt)}.\exp(-\frac{1}{RT}(V_f(mut) - V_f(wt))) \tag{7.4}$$

The benefit of V_f is that it can be directly computed by molecular simulation and subsequently used to assess the enzymatic fitness of a mutant under chemotherapeutic pressure. Furthermore, using Equation 7.4, it can be directly transformed into the *vitality* measure in the regime where mutants do not cause significant changes in values of k_{cat} ($(k_{cat})_{mut}/(k_{cat})_{wt} \sim 1$). Note that, as in the study reported by Gulnik *et al.* [224], the vitality of the wildtype is necessarily unity whilst the relative free energy potential ($\Delta V_f = V_f(mut) - V_f(wt)$) is zero. Mutants which increase the relative fitness potential ($\Delta V_f \geq 0$) will confer no advantage over the wildtype, whilst those with $\Delta V_f < 0$ will be advantageous under drug pressure. In this way the path of increased enzymatic fitness tends towards a minimum in the mutational landscape of the fitness potential.

7.3 Methods

The implementation of the substrate-bound HIV-1 protease study was almost identical in method to that described for the saquinavir-bound protease study described in § 6.2 and made use of elements of the Binding Affinity Calculator described in Chapter 6 and in Appendix A. We will describe differences in the protocol here.

7.3.1 Initial Preparation of Models

Several structures of D25N inactive HIV-1 protease have recently been reported bound to a variety of the ten recognised cleavage sites (see Chapter 3) [128, 129, 131]. The 1TSU structure (2.1 Å resolution), containing the protease bound to the NC-p1 substrate, was used as the starting point for all substrate simulations. The NC-p1 substrate consists of an octa-peptide chain, with amino acid sequence RQAN-FLGK, where the peptide bond cleaved by the protease is in between the asparagine and phenylalanine

residues. When bound to the protease, the peptide chain residues are labelled the P4-P4′ subsites respectively. The NC-p1 substrate was selected instead of the other available structures as it is one of the candidates for the rate determining step of the cleavage process [229] and because it has shown to be susceptible to Gag mutations in response to drug therapy [230]. The ff03 forcefield [24] was sufficient to describe all protein and substrate parameters. In order to obtain an active protease, the N25D mutation was incorporated on the catalytic dyad for all substrate systems. These mutations, as well as the G48V and L90M mutations were implemented using VMD.

When considering the substrate, the favoured mechanism for catalysis involves a monoprotonated dyad in which the aspartic acid that is protonated is coordinated next to the carbonyl oxygen adjacent to the peptide bond to be cleaved [149]. However, in the substrate crystal structure the corresponding aspartic acid belongs to the second monomeric chain. We therefore reversed the designation of the two chains in the crystal structure and assigned monoprotonation to the newly designated D25 residue.

The Leap module [248] in the AMBER 9 software package [53] was then used to combine each apo-protease system with the ligand. Nine Cl^- counter-ions were added to electrically neutralise each substrate-bound system. Each system was then solvated using atomistic TIP3P water [249] in a cubic box with at least 14 Å distance around the complex. The size of each substrate-bound system was 41332, 41350, 41328 and 41346 atoms for the wildtype, G48V, L90M and G48V/L90M systems respectively.

7.3.2 Minimisation, Equilibration, Production and Free Energy Protocols

The minimisation, equilibration and production protocols implemented in the simulation of all substrate-bound HIV-1 proteases were exactly the same as those described in Chapter 6. Furthermore, the MMPBSA and normal mode analysis procedures were also identical to those described in Chapter 6.

7.4 Results and Discussion

7.4.1 Structural Flexibility of Monoprotonated HIV-1 Protease/NC-p1 Substrate Complexes

All four proteases moved a similar amount from the initial crystal structure and exhibited almost identical backbone flexibilities. Root mean squared deviations (RMSDs) relative to the starting crystal structure of 1.00 ± 0.09 Å, 1.25 ± 0.10 Å, 1.05 ± 0.09 Å and 1.05 ± 0.09 Å were exhibited for the wildtype, G48V, L90M and G48V/L90M mutants respectively in the production phase.

The substrate extended from subsite P4 to P4′, the end subsites being completely immersed in the solvent (see Figure 7.1(a)). We assessed the flexibility of each of the subsites of the NC-p1 substrate in all protease systems by determining the root mean squared fluctuations (RMSF) for all non-hydrogen atoms belonging to each residue (see Figure 7.1(b)(i)) and for just the backbone atoms (see

Figure 7.1(b)(ii)). The calculation was performed over the entire 10 ns trajectory; however as subsites P4 and P4′ were exposed to solvent and exhibited substantial flexibility upon visual inspection, the average structure used in determining the RMSF was obtained through prior alignment of only the P3-P3′ backbone.

In general, the backbone flexibility between the P3 and P3′ subsites was both small, with RMSFs less than 0.5 Å, and invariant across all proteases. When incorporating the flexibility of the side-chains, the L90M exhibited a significant increase in the flexibility of the P1′ subsite over the other systems at this position, marked by an increased RMSF of 0.5 Å. Interestingly, there was significantly increased RMSF for the P4′ subsite in each system as compared to the other subsites. P4′ was composed of a lysine residue which visual inspection confirmed was orientated directly into the solvent. During the course of the simulation, large conformational reorientations of the residue were observed in response to solvent interaction which were pivoted around its backbone atoms. This explained the large increase in RMSF observed for this residue. By contrast, even though the P4 subsite which was composed of arginine was also exposed to the solvent, strong interactions were observed between it and several protease residues. This resulted in reduced motion of the residue in response to solvent interactions and explained the subsequently reduced RMSF observed for this residue.

We also compared the total RMSD of the NC-p1 substrate (R_t) with the RMSD induced through only conformational changes (R_{cc}) in the substrate (see Figure 7.2). As in the studies reported in Chapters 5 and 6, protease backbones were aligned prior to determining R_t, whilst substrate backbones were aligned prior to determining R_{cc}. The NC-p1 substrate exhibited ~3 Å conformational motion away from the starting structure in all systems. However, whilst the total RMSD of the substrate in the G48V and L90M systems was only ~4 Å by the end of the simulation, for the wildtype and G48V/L90M mutant, it was around 6 Å and 5 Å respectively. The increased separation of R_t from R_{cc} in these two systems, indicated a larger degree of rotational and/or translational motion of the centre of mass of the substrate as compared to the G48V and L90M mutants.

Interestingly, in the 1TSU crystal structure of NC-p1 bound to the protease the P4 subsite, which is composed of arginine, is orientated into the hydrophobic core of the S2 pocket of the protease, formed by residues V132, I184 and I147 as well as the hydrophilic residues D129 and D130. Additionally, the P4′ subsite, composed of lysine, is orientated into the S4′ subsite of the protease, composed of residues D30, K45, M46 and I47. The significant value of R_t observed in all systems is predominantly explained by the high mobility of the solvent-exposed P4′ subsite (see Figure 7.1(b)), which in all systems exhibited large conformational changes away from the S4′ subsite. Furthermore, visual inspection confirmed this P4′-motion to span the entire entrance to the active site.

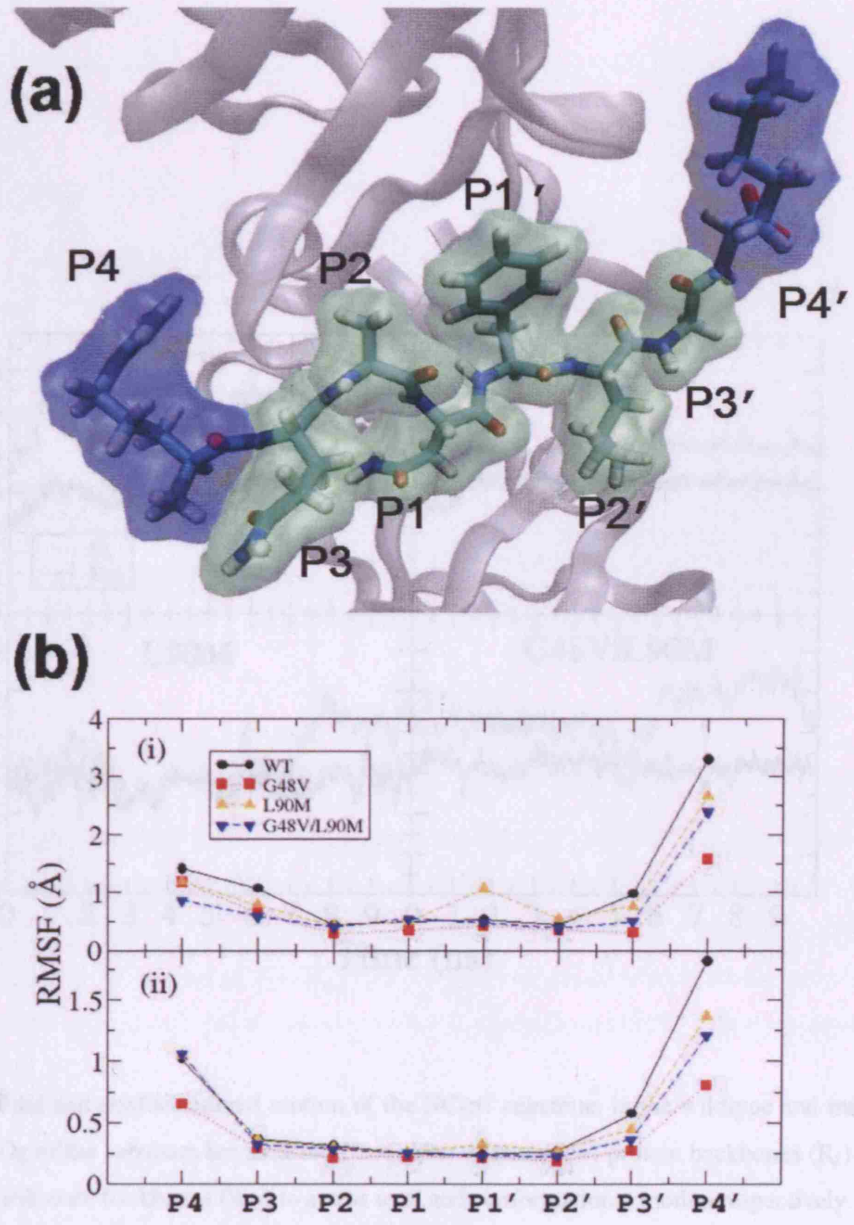


Figure 7.1: Subsite flexibility of the NC-p1 substrate for wildtype HIV-1 protease and the G48V, L90M and G48V/L90M mutants. (a) Schematic of the NC-p1 substrate bound to the wildtype protease (top down view: flaps removed for clarity). Subsites P4 and P4′ (blue) are fully exposed to the solvent whilst P3 to P3′ (green) lie substantially within the active site. (b) Subsite decomposition of the RMSF calculated over all 10 ns of production, across (i) all non-hydrogen atoms of the substrate and (ii) across just the backbone atoms of the substrate.

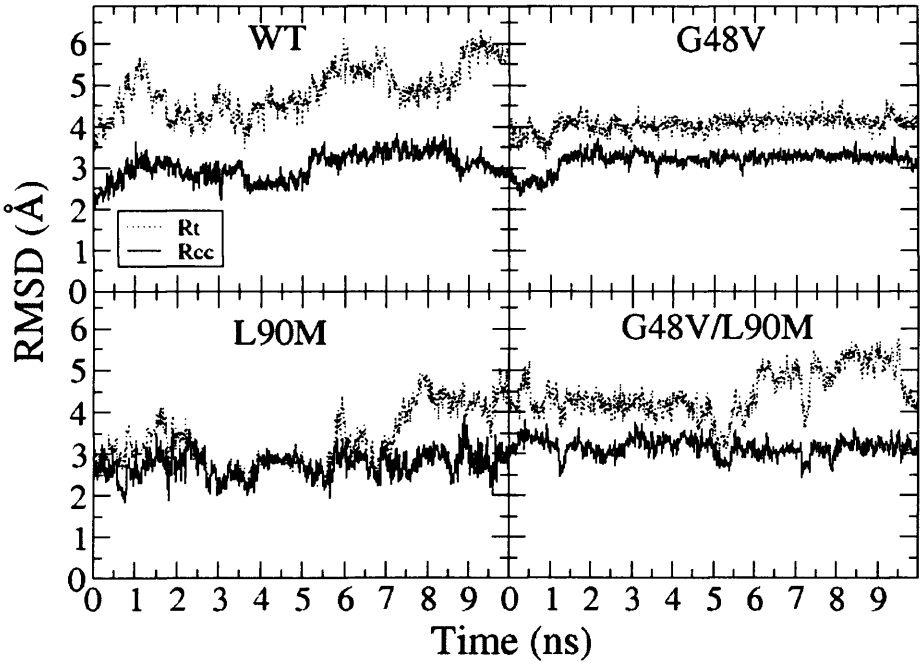


Figure 7.2: Total and conformational motion of the NC-p1 substrate, in the wildtype and mutant proteases. RMSDs of the substrate are measured both after alignment of protein backbones (R_t) and after alignment of substrate backbones (R_{cc}) to assess total and conformational motion respectively. The NC-p1 substrate exhibited substantially more conformational motion than saquinavir (see Chapter 6), with R_{cc} around 3 Å in all systems. The wildtype and G48V/L90M systems also exhibited significant total motion of the substrate away from the original crystal structure, up to ~6 Å and ~5 Å respectively.

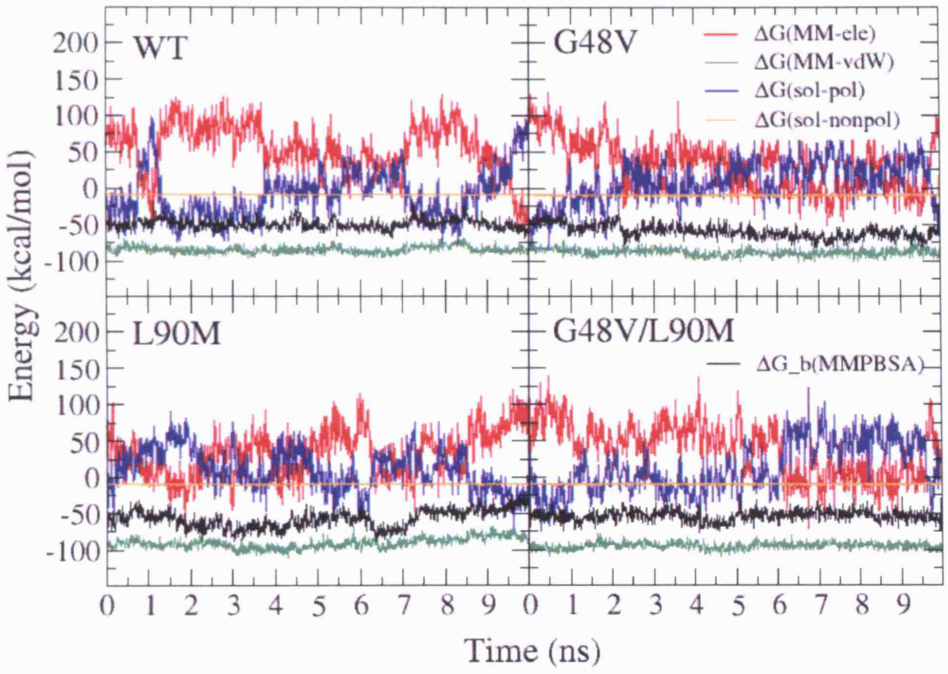


Figure 7.3: Time evolution of each independent component of the MMPBSA contribution to the free energy difference of substrate-binding across 1000 snapshots of the 10 ns production run for each substrate-protease system. The components of the free energy exhibited large fluctuations over the entire time interval. The variation of ΔG_{ele}^{MM} and ΔG_{ele}^{sol} is as large as $\sim$100 kcal/mol; even though these two components are anti-correlated and result in stable fluctuations of ΔG_{ele}^{MM}, fluctuations were still as large as $\sim$12 kcal/mol.

7.4.2 Time-Series and Convergence Analysis of the Enthalpic and Entropic Components of Substrate-Binding

We determined convergence properties by analysing the time evolution of each of the components of the free energy for all substrate-protease systems over the 10 ns trajectories (see Figure 7.3). In all systems, both electrostatic components, ΔG_{ele}^{MM} and ΔG_{ele}^{sol} exhibited substantial variation ranging over $\sim$100 kcal/mol. This pair of components were significantly anti-correlated with cross-correlation coefficients of -0.98, -0.97, -0.95 and -0.98 for the wildtype, G48V, L90M and G48V/L90M mutant systems. Interestingly, the van der Waals component, ΔG_{ele}^{vdW}, also exhibited significant fluctuations (up to $\sim$7 kcal/mol). This led to the overall enthalpic contribution, G_b^{MMPBSA}, exhibiting stable but still considerably large fluctuations (up to $\sim$12 kcal/mol).

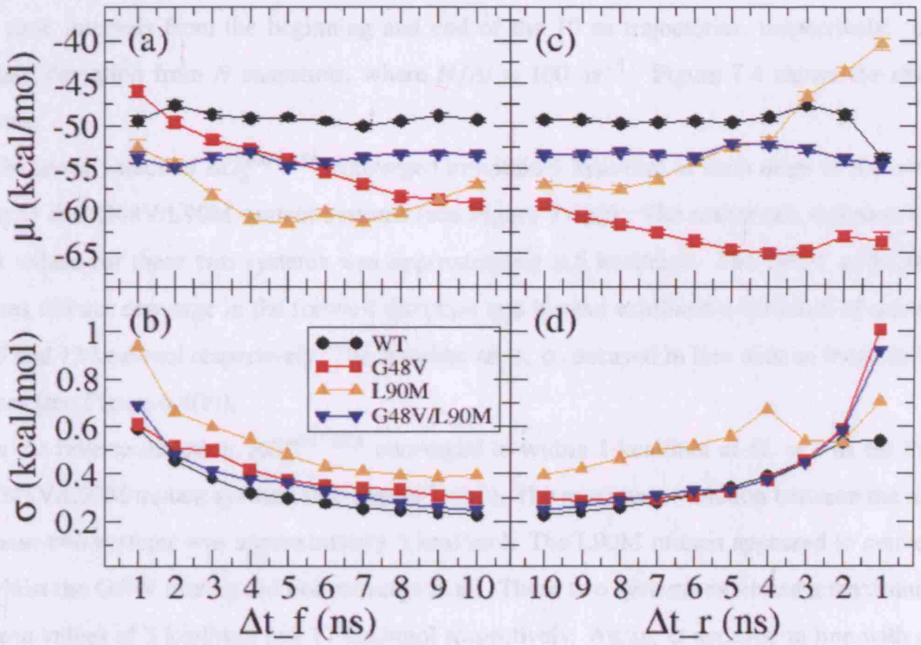


Figure 7.4: Convergence of the enthalpic component of binding, ΔG_b^{MMPBSA}, assessed by (a) the mean (μ) and (b) the standard error ($\sigma = \sigma_{sd}/N^{1/2}$) of the forward ($\Delta t_f$) and (c),(d) reverse ($\Delta t_r$) time intervals across the 10 ns trajectories for each substrate-protease system. σ_{sd} is the standard deviation and N, the number of snapshots, where $N/\Delta t = 100\ \text{ns}^{-1}$. The mean value of ΔG_b^{MMPBSA} converged for the wildtype and G48V/L90M mutant systems within 1 kcal/mol at Δt_f = 6 ns, whilst in reverse, convergence to within a similar threshold occurred at Δt_r = 7 ns. Convergence of the L90M mutant system seemed to occur only in the reverse direction after 8 ns and did not occur at all in the G48V mutant system. The standard error for all systems decayed, as expected, with increased N, except for the L90M mutant, which exhibited increased fluctuations at $\Delta t_r = 4$ ns.

The substantial flexibility of the NC-p1 protease in the HIV-1 protease active site (see § 7.4.1) was reflected in the large fluctuations exhibited in the MMPBSA calculations over the 10 ns time period. This was indicative that, unlike the case of drug-bound proteases, convergence of the time-averaged components of the free energy may need a larger time-scale than the 10 ns studied here.

Nonetheless, we assessed the convergence of the enthalpic contribution to the free energy of binding, ΔG_b^{MMPBSA}, analogous to the method employed for the drug-bound proteases. We calculated the mean (μ) and standard error ($\sigma = \sigma_{sd}/N^{1/2}$) of ΔG_b^{MMPBSA} as a function of both the forward (Δt_f) and reverse (Δt_r) time intervals from the beginning and end of the 10 ns trajectories, respectively. σ_{sd} was the standard deviation from N snapshots, where $N/\Delta t = 100$ ns^{-1}. Figure 7.4 shows the results of this analysis.

The mean values of ΔG_b^{MMPBSA} converged to within 1 kcal/mol of each other at $\Delta t_f = 6$ ns for the wildtype and G48V/L90M mutant systems (see Figure 7.4(a)). The maximum variation between the mean values for these two systems was approximately 2.5 kcal/mol. The G48V and L90M mutant systems did not converge in the forward direction and instead exhibited a variation of mean values up to 8.5 and 13 kcal/mol respectively. The standard error, σ, decayed in line with an increase in snapshot number (see Figure 6.4(b)).

In the reverse direction, ΔG_b^{MMPBSA} converged to within 1 kcal/mol at $\Delta t_r = 7$ ns for the wildtype and G48V/L90M mutant systems (see Figure 7.4(c)). The maximum variation between the mean values for these two systems was approximately 5 kcal/mol. The L90M mutant appeared to converge after 8 ns, whilst the G48V mutant did not converge at all. These two systems exhibited a maximum variation of mean values of 5 kcal/mol and 17 kcal/mol respectively. Again, σ decayed in line with an increase in snapshot number (see Figure 7.4(d)), except for $\Delta t_r = 4$ ns in the L90M system, which exhibited increased fluctuations.

The variation of the components of the configurational entropy contribution for all systems, assessed by normal mode analysis, are shown in Figure 7.5. Whilst the translational ($T\Delta S_{tra}$) and rotational ($T\Delta S_{rot}$) components of the entropy were well-behaved and effectively constant, there was substantial variation in the vibrational ($T\Delta S_{vib}$) component of the entropy across the different snapshots, with a range of $\sim$ 45 kcal/mol and a standard deviation of approximately 8 kcal/mol.

The convergence of the entropic component of binding, $T\Delta S$, was determined in the same way to that described for the enthalpy (see Figure 7.6) with the exception that the number of snapshots used was $N = 50$ across the 10 ns trajectory ($N/\Delta t = 5$ ns^{-1}). The mean values of $T\Delta S$ within a system converged to within 0.5 kcal/mol of each other at $\Delta t_f = 8$ ns for all systems (see Figure 7.6(a)). There was as much as 10 kcal/mol variation in the mean value between the first nanosecond and the converged value, which occurred in the L90M system. The standard error, σ, for the first 6 ns deviated from an expected decay (see Figure 7.6(b)), followed by reversion to decay with an increase in snapshot number.

In the reverse direction, $T\Delta S$ also converged to within 0.5 kcal/mol at $\Delta t_r = 9$ ns for all systems (see

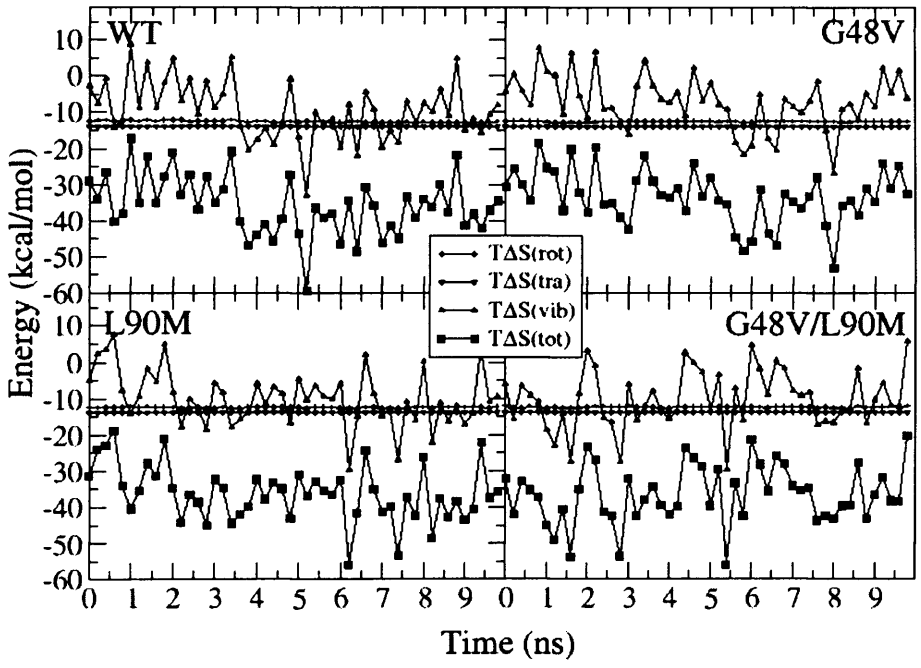


Figure 7.5: Time evolution of the components of configurational entropy, $T\Delta S_{rot}$, for all substrate-protease systems. 50 equally spaced snapshots were selected across the 10 ns trajectory. The values of $T\Delta S_{rot}$ and $T\Delta S_{tra}$ were effectively constant across all trajectories. Large variation with a range of ~ 45 kcal/mol was observed for $T\Delta S_{vib}$ as well as fluctuations with a standard deviation of approximately 8 kcal/mol.

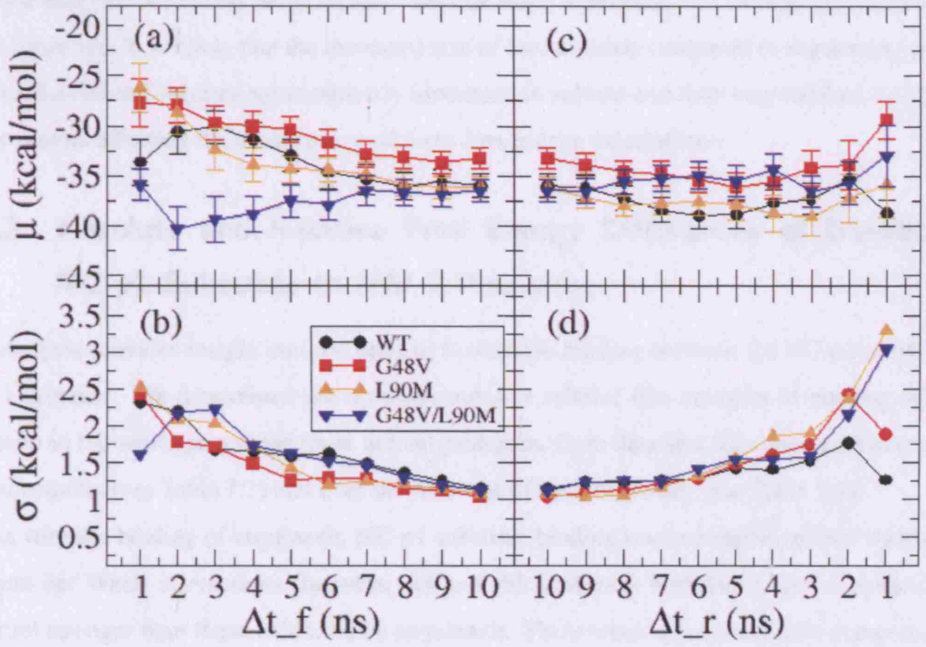


Figure 7.6: Convergence of the entropic component of binding, $T\Delta S$, assessed by (a) the mean (μ) and (b) the standard error ($\sigma = \sigma_{sd}/N^{1/2}$) of the forward ($\Delta t_f$) and (c),(d) reverse ($\Delta t_r$) time intervals across the 10 ns trajectories for each substrate-protease system. σ_{sd} is the standard deviation and N, the number of snapshots, where $N/\Delta t = 5$ ns^{-1}. The mean value of $T\Delta S$ converged for all systems within 0.5 kcal/mol at $\Delta t_f = 8$ ns; in reverse, convergence to below a similar threshold occurred at $\Delta t_r = 9$ ns. For the first 6 ns and 4 ns in the forward and reverse directions respectively, deviation was exhibited from the expected decay of the standard error σ, followed by reversion to an expected decay with increased N.

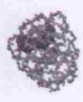


Figure 7.6(c)), whilst the maximum variation between the last nanosecond and the first was approximately 5 kcal/mol. Similar deviations were observed for the decay of σ in the initial 4 ns, followed by normal decay with an increase in N (see Figure 7.6(d)).

Convergence analysis revealed that only two of the four substrate protease complexes studied here had converged after 10 ns of production, namely the wildtype and G48V/L90M mutant systems. The lack of convergence exhibited by the other two systems and the increased fluctuations in all substrate-bound systems as compared to drug-bound systems, highlights the difficulty in attaining accurate absolute free energies of binding of the NC-p1 substrate using MMPBSA and configurational entropy over a 10 ns timescale. It is likely that the increased size of the substrate compared to saquinavir, coupled with the fact that its end subsites are completely immersed in solvent and thus very flexible (see § 7.4.1), are major factors affecting the convergence of these free energy calculations.

7.4.3 Absolute and Relative Free Energy Differences of Binding of the NC-p1 Substrate to HIV-1 Proteases

Our analysis provides insight into the basis of favourable binding between the NC-p1 substrate and the HIV-1 protease. We determined the absolute and thus relative free energies of binding of the NC-p1 substrate to the wildtype and the three mutant proteases, from data sets time-averaged over all 10 ns of the trajectories (see Table 7.1) and over the first 1 ns of each trajectory (see Table 7.2).

As with the binding of saquinavir, NC-p1 substrate binding was principally driven by highly attractive van der Waals interactions (between -84 and -95 kcal/mol) with the protease, approximately 15 kcal/mol stronger than those exhibited by saquinavir. These were only moderately compensated for by an opposing repulsive total electrostatic contribution to binding, ranging from 39 to 52 kcal/mol. Unlike for saquinavir however, the gas-phase electrostatic interaction, ΔG_{ele}^{MM}, was significantly repulsive (32 to 54 kcal/mol), whilst the polar solvation component, ΔG_{pol}^{sol}, was only slightly repulsive ($\sim$8-13 kcal/mol), except for the wildtype, which exhibited a slightly attractive ΔG_{pol}^{sol} (-8.12 kcal/mol).

Previous experimental studies determined the Michaelis constant, K_M, for the NC-p1 substrate as 0.17 mM for wildtype protease [230], which corresponds to an absolute free energy difference of binding of -5.35 kcal/mol at 310.15 K. The substrate therefore binds much less strongly than saquinavir, for which nanomolar potency is exhibited and with a subsequent 9.11 kcal/mol difference in binding. In our study, the inclusion of configurational entropy again greatly improved the value of the absolute free energy of binding to -13.42 kcal/mol as opposed to using just MMPBSA alone, for which a value of -49.32 kcal/mol was obtained. Nonetheless, our method still substantially overestimates the binding of the substrate to the wildtype by 8.3 kcal/mol as compared to experiment, unlike for saquinavir, in which the maximum deviation away from the experimental value was only $\sim$1 kcal/mol. It is likely that this overestimate is due to the limitations of the single-trajectory method. The NC-p1 substrate is

Enthalpic components of binding ($\Delta t_{f/r}$ = 10 ns)

Sequence	ΔG^{MM}_{vdW}	ΔG^{MM}_{ele}	ΔG^{sol}_{nonpol}	ΔG^{sol}_{pol}	ΔG^{tot}_{ele}	ΔG^{MMPBSA}_{b}
WT	**-84.71** (0.153)	**53.16** (1.122)	**-9.65** (0.009)	**-8.12** (1.069)	**45.04** (0.223)	**-49.32** (0.227)
G48V	**-88.17** (0.166)	**30.84** (1.072)	**-10.55** (0.007)	**8.46** (0.891)	**39.30** (0.291)	**-59.42** (0.292)
L90M	**-90.86** (0.228)	**32.79** (1.007)	**-9.95** (0.009)	**10.99** (0.880)	**43.78** (0.313)	**-57.03** (0.395)
G48V/L90M	**-94.93** (0.145)	**37.91** (1.135)	**-9.84** (0.007)	**13.32** (1.103)	**51.23** (0.230)	**-53.54** (0.243)

Entropic components and absolute free energy differences of binding ($\Delta t_{f/r}$ = 10 ns)

Sequence	$T\Delta S_{tra}$	$T\Delta S_{rot}$	$T\Delta S_{vib}$	$T\Delta S_{tot}$	ΔG_b
WT	**-13.87** (0.000)	**-12.55** (0.028)	**-9.48** (1.144)	**-35.90** (1.158)	**-13.42** (1.385)
G48V	**-13.87** (0.000)	**-12.54** (0.014)	**-6.91** (1.055)	**-33.32** (1.059)	**-26.10** (1.351)
L90M	**-13.87** (0.000)	**-12.47** (0.017)	**-9.86** (1.086)	**-36.20** (1.092)	**-20.83** (1.487)
G48V/L90M	**-13.87** (0.000)	**-12.40** (0.020)	**-10.25** (1.137)	**-36.52** (1.144)	**-17.02** (1.387)

Mean energies are in kcal/mol, corresponding standard errors in parentheses.

Enthalpic sample size: $N/\Delta t = 100\ \text{ns}^{-1}$, entropic sample size: $N/\Delta t = 5\ \text{ns}^{-1}$

Table 7.1: Enthalpic and entropic decomposition of the NC-p1 substrate binding to HIV-1 protease wildtype and mutants time-averaged over all 10 ns.

Enthalpic components of binding (Δt_f = 1 ns)

Sequence	ΔG^{MM}_{vdW}	ΔG^{MM}_{ele}	ΔG^{sol}_{nonpol}	ΔG^{sol}_{pol}	ΔG^{tot}_{ele}	ΔG^{MMPBSA}_{b}
WT	**-83.13** (0.488)	**55.73** (2.920)	**-9.60** (0.029)	**-12.45** (2.977)	**43.28** (0.688)	**-49.46** (0.592)
G48V	**-83.40** (0.423)	**82.24** (2.296)	**-10.49** (0.021)	**-34.34** (2.103)	**47.90** (0.697)	**-45.99** (0.612)
L90M	**-90.14** (0.479)	**30.30** (2.277)	**-9.85** (0.026)	**16.97** (2.227)	**47.28** (0.805)	**-52.71** (0.941)
G48V/L90M	**-97.54** (0.403)	**80.36** (2.351)	**-9.95** (0.015)	**-26.75** (2.028)	**53.61** (0.681)	**-53.88** (0.700)

Entropic components and absolute free energy differences of binding (Δt_f = 1 ns)

Sequence	$T\Delta S_{tra}$	$T\Delta S_{rot}$	$T\Delta S_{vib}$	$T\Delta S_{tot}$	ΔG_b
WT	**-13.87** (0.000)	**-12.32** (0.027)	**-6.94** (1.210)	**-33.12** (1.216)	**-16.34** (1.808)
G48V	**-13.87** (0.000)	**-12.46** (0.029)	**-3.95** (1.301)	**-30.28** (1.310)	**-15.71** (1.922)
L90M	**-13.87** (0.000)	**-12.47** (0.027)	**-3.34** (1.800)	**-29.68** (1.800)	**-23.03** (2.741)
G48V/L90M	**-13.87** (0.000)	**-12.43** (0.040)	**-10.68** (1.585)	**-36.98** (1.605)	**-16.90** (2.305)

Mean energies are in kcal/mol, corresponding standard errors in parentheses.

Enthalpic sample size: $N/\Delta t = 100\ \text{ns}^{-1}$, entropic sample size: $N/\Delta t = 20\ \text{ns}^{-1}$

Table 7.2: Enthalpic and entropic decomposition of the NC-p1 substrate binding to HIV-1 protease wildtype and mutants time-averaged over all 1 ns.

larger and more flexible than saquinavir and it is plausible that in the unbound state it exhibits much more conformational freedom than that described by a complex-extracted structure. Using the three-trajectory method to account for the significantly increased loss of configurational entropy upon binding that would result from this would likely help to narrow the difference between the experimental value and that attained from the single-trajectory method used here.

Interestingly, the gas-phase van der Waals and non-polar solvation interactions compared well with that of Wang and Kollman [218]. However, the overall electrostatic penalty exhibited here was less severe by ~40 kcal/mol and resulted in a far larger total binding enthalpy than the -15.4 kcal/mol observed by Wang and Kollman [218]. This discrepancy partially arises from the difference in the chosen substrates; the CA-p2 substrate used in that study has a different amino acid composition (ARVL-AEAM) than NC-p1 (RQAN-FLGK), and it is natural to assume it will have different binding energetics. This is further supported by the variation in observed catalytic efficiencies across the natural substrates [229]. However, it is also likely that in our study, the total electrostatic penalty, which exhibited large fluctuations, was underestimated and this may explain the overestimation of the total binding as compared with experiment. Furthermore, as the entropic contribution was not included in the study by Wang and Kollman [218], and from our studies was shown to be highly significant in determining an accurate absolute value of binding, it is likely that Wang and Kollman [218] significantly underestimated the binding enthalpy.

Due to the convergence problems, discussed in § 7.4.2, it was only possible to accurately compare the relative binding free energies of the wildtype and G48V/L90M mutant (see Table 7.1). The G48V/L90M mutant exhibited ~4 kcal/mol increased binding enthalpy moderated by only a ~0.5 kcal/mol decreased configurational entropy over the wildtype, leading to a ~3 kcal/mol increased binding free energy. Interestingly the enthalpy obtained in the first 1 ns for these two systems (see Table 7.1) was almost identical to that obtained in the 10 ns time-average. Entropic values for the wildtype and G48V/L90M systems differed by ~2.8 kcal/mol and ~0.5 kcal/mol respectively across these two time-averaged sets.

Although a complete description of the catalytic efficiency of the protease is provided by the specificity constant k_{cat}/K_m (see Chapter 1), increases in binding affinity of the substrate correspond directly to a decrease in K_m, and are thus indicative of increased catalytic efficiency. Based on this, our results support the notion that the G48V/L90M mutant substantially increases the catalytic efficiency of the protease as compared to the wildtype. Unfortunately, there are no experimental results for the binding of the NC-p1 substrate with the G48V/L90M mutant. However, the L90M mutant has been shown to enhance the catalytic efficiency by 2.5 fold over the wildtype for the NC-p1 substrate [230]. If k_{cat} is assumed not to change, then the upper limit of the increase in binding from these experiments is approximately 0.6 kcal/mol. Unfortunately, for this mutant, our results were not able to provide a converged free energy value for comparison.

However, by combining the most appropriate results of the G48V and L90M systems we are able

to provide a more qualitative comparison of the 4 systems. Considering the consistency of enthalpic values in the wildtype and G48V/L90M systems in between the 1 ns and the 10 ns time-averages and noticing that considerable drift in the enthalpy was exhibited in G48V and L90M, we selected the 1 ns time-averaged enthalpies of these two mutants for further comparison. This was substantiated by the fact that the large changes in enthalpy (~15 kcal/mol) across the 10 ns production phase are physically unreasonable (see Figure 7.4). It is then likely that the 1 ns time-averages for these two mutants correspond to more accurate enthalpies, as they are determined before considerable drift of the enthalpy has taken place. Additionally, the G48V and L90M enthalpies in the 1 ns time-averages are within ~3.5 kcal/mol of the wildtype, whilst an unrealistic deviation of ~10 kcal/mol and ~8 kcal/mol are exhibited respectively in the 10 ns time-averaged trajectories. Conversely, as the entropy in all systems did converge, it is more appropriate to assign the 10 ns time-average results for the G48V and L90M systems.

Sequence	ΔG_b^{MMPBSA}	$T\Delta S_{tot}$	ΔG_b	$\Delta G(exp)$
WT	**-49.32** (0.227)	**-35.90** (1.158)	**-13.42** (1.385)	**-5.35** (NA)
G48V	**-45.99*** (0.612)	**-33.32** (1.059)	**-12.67** (1.671)	NA
L90M	**-52.71*** (0.941)	**-36.20** (1.092)	**-16.51** (2.033)	**-5.91**** (NA)
G48V/L90M	**-53.54** (0.243)	**-36.52** (1.144)	**-17.02** (1.387)	NA

Mean energies are in kcal/mol, corresponding standard errors in parentheses.

Values taken from 10 ns time-averages except where mentioned. * Taken from 1 ns time-average $\Delta G(exp)$ calculated from K_m [230] at 310.15 K. NA: Not available. ** Assumes relative $k_{cat} = 1$

Table 7.3: Binding free energies of the NC-p1 substrate to HIV-1 protease wildtype and mutants.

Combining these results, we obtained a refined set of binding free energies for the NC-p1 substrate to the four proteases (see Table 7.3). These results imply that the G48V mutant marginally decreases catalytic efficiency whilst the L90M mutant enhances it. The ~3 kcal/mol increase in binding for the L90M mutant agrees qualitatively with the possible 0.6 kcal/mol exhibited in experiment. However, as experimental values for the change in k_{cat} and K_m for the L90M mutant were not reported separately in the study by Feher *et al.* [230], it is not possible to determine whether our calculation overestimates the change in binding or whether our results imply a substantial decrease in K_m, which is compensated for partially by a reduction in k_{cat}. In the latter case, the increase in experimental binding free energy would then be considerably larger than the 0.6 kcal/mol specified here.

7.4.4 Ranking of Enzymatic Fitness

We combined the values for the free energy of binding of saquinavir, reported in Chapter 6, with the binding free energies of the substrate to construct the free energy potential of enzymatic fitness, $V_f(\mu)$

and the corresponding approximate vitality metrics (V) of each of the mutants (see Table 7.4).

Sequence	ΔG_{sub}	ΔG_{drug}	$V_f(\mu)$	$\Delta V_f(\mu)$	V
WT	**-13.42** (1.385)	**-13.76** (1.660)	**0.34** (3.045)	**0**	1
G48V	**-12.67** (1.671)	**-11.31** (1.264)	**-1.36** (2.935)	**-1.70**	15.7
L90M	**-16.51** (2.033)	**-11.55** (1.527)	**-4.96** (3.560)	**-5.30**	5.4×10^3
G48V/L90M	**-17.02** (1.387)	**-10.64** (1.459)	**-6.38** (2.846)	**-6.72**	5.4×10^4

Energy terms are in kcal/mol. The *vitality* metric is dimensionless.

ΔG_{drug} values taken from Δt_r = 4 ns time-averaged trajectories in Chapter 6.

Table 7.4: Vitality metric, V, and enzymatic fitness potential, $V_f(\mu)$.

Due to the lack of convergence in the G48V and L90M results, care has to be taken when interpreting the corresponding values of the enzymatic fitness potential. However, based on our analysis, both the G48V and L90M mutants exhibit a substantial decrease in the fitness potential, corresponding to an increase in enzymatic fitness over the wildtype. This agrees qualitatively with the increase in vitality for these mutants for different Gag substrates [15]. Unfortunately, in the study by Maschera *et al.* [15], the NC-p1 was not studied, so a direct comparison is not possible, although increased vitality was exhibited for all Gag substrates. The convergence of the G48V/L90M mutant free energies allowed a more quantitative comparison with the wildtype. The fitness potential decreased by 6.72 kcal/mol with respect to the wildtype; not only does the G48V/L90M mutant increase drug resistance it also increases the catalytic efficiency of the NC-p1 substrate. These two effects are complementary and increase the overall fitness of the mutant with respect to the wildtype.

The energetic differences between inhibitor and substrate binding and their combination in terms of a fitness potential are able to provide molecular insight into the generally observed order of mutations *in vivo*. Our results favour the accumulation of L90M prior to G48V in the resistance pathway, due to it exhibiting a lower fitness potential. The increase in vitality and decrease in the free energy potential by the L90M mutation and the even greater vitality increase shown by the G48V/L90M double mutation are consistent with the observed pattern of emergence of these mutations *in vivo* [269] as well as experimental measurements of enzymatic fitness [228].

The L90M mutation, which appears quickly in response to treatment with several inhibitors and only marginally reduces the binding affinity of saquinavir, may be selected because it provides an effective platform for other mutations. This is effected by increasing enzymatic fitness through the improvement in catalytic efficiency whilst under chemotherapeutic pressure. The G48V mutation when in concert with L90M further increases fitness by significantly altering drug binding, whilst also increasing catalytic efficiency. However, even though the G48V mutation alone confers more drug resistance than L90M, its overall enzymatic fitness is less, due to the increased catalytic efficiency conferred by L90M

over both the wildtype and the G48V mutant.

A free-energy based metric for the description of enzymatic fitness provides a natural way of mapping out the fitness landscape of HIV-1 protease in response to a specific inhibitor. In this study, we have only considered one inhibitor and three drug resistant mutants. Furthermore, we have made the assumption that k_{cat} does not vary significantly from wildtype to mutant, in line with previous experimental work on a range of mutants [224]. Whilst such an assumption is valid for the mutations considered here [15], it may not apply to some drug resistant mutations. The NC-p1 substrate has been used in our simulation as it is relatively slowly cleaved [229]. The CA-p2 substrate is also slowly cleaved and so both are candidates for the rate determining step of viral maturation. Our computation therefore also assumes that a change in the binding properties of the NC-p1 substrate alone has an overriding bearing on the enzymatic fitness and ignores the corresponding effects of mutations on CA-p2 processing. This is a plausible assumption, given that Gag cleavage site mutations which enhance catalytic efficiency are selected in the NC-p1 substrate [230, 237] and not in CA-p2, indicating that enhancement of NC-p1/protease binding dominates the resulting increase in enzymatic fitness under drug pressure.

7.5 Conclusion

A complete description of drug resistance must incorporate the effects of a mutation on catalytic efficiency. Using the recently reported [131] crystal structure of inactive HIV-1 protease bound to the rate limiting NC-p1 substrate, we studied the effect of the G48V, L90M and G48V/L90M mutations on substrate binding. We calculated relative and absolute binding affinities for the NC-p1 substrate using molecular dynamics simulations over a production phase of 10 ns.

Unlike the case of inhibitor binding, reported in Chapter 6, a 10 ns duration was not sufficient to guarantee convergence of all free energies calculated. This was likely due to the increased size and flexibility of the substrate over the inhibitor, as well as complete exposure of terminal subsites to the solvent. We suggest two strategies for overcoming such convergence problems. Firstly, longer simulations may allow non-converged systems to stabilise and secondly, modifying the starting structure by removing the terminal P4 and P4′ subsites would significantly reduce the flexibility of the substrate allowing convergence in a shorter period of time.

Based on the convergence of the wildtype and G48V/L90M mutant, our results show that the G48V mutation in concert with L90M increases substrate binding as compared to the wildtype. This is complementary to the effect of the mutant on inhibitor binding and further causes an increase in the overall enzymatic fitness of the mutant over the wildtype.

We have additionally developed the free energy potential of enzymatic fitness (V_f), based on the vitality metric devised by Gulnik *et al.* [224] as a quantitative way of gauging enzymatic fitness from molecular simulation. Although, we have only been able to explore it qualitatively, due to the lack

of convergence in our simulations, in principle it can be used to provide insight into the pathways of observed resistance *in vivo*.

CHAPTER 8

Conclusions and Future Directions

HIV-1 protease is one of the key targets for anti-retroviral inhibitors, designed to treat people infected with HIV. The development of drug resistant mutations in HIV-1 protease in response to therapy is well known, often reducing the strength of binding of inhibitors to the protease. Whilst well established experimental techniques exist to determine the strengths of molecular association, namely the binding affinity, between protease and inhibitor, it nonetheless still remains very difficult to elucidate the molecular basis of the resistance conferred by mutations. In this thesis we have investigated the molecular basis of drug resistance in HIV-1 protease conferred by mutants that emerge in response to treatment with the inhibitor saquinavir. We have used the well established computational technique of molecular dynamics (MD) to probe both the molecular mechanisms of resistance as well as the thermodynamic basis for the changes in experimentally observed binding affinities.

With the benefit of high performance computing (HPC) and grid technology we have been able to conduct both multiple and long, fully atomistic MD simulations for a range of ligand-protease variants. Such extended simulations provide insight into a mutation-assisted lateral escape mechanism for saquinavir from the active site of the protease. We have discussed how this mechanism alters the conventional view of protease-inhibitor dissociation, from that of a 'fully-open' protease to one which need only be 'semi-open'. Furthermore, we discuss how the fast timescale of such an event may allow mutations an alternative kinetic mechanism with which to confer resistance.

Molecular mechanisms are difficult to determine experimentally. Computational studies, on the other hand, are limited by the wall-clock time required to compute a sufficient amount of simulated time. We have independently utilised steered molecular dynamics (SMD) and principal component analysis (PCA) methods to investigate the basis of the lateral dissociation mechanism reported here. Interestingly, a recent method combined PCA and SMD to correctly direct the steering of a peptide-folding simulation in a direction governed by the principal components of the system [270]. Such a synergistic approach may allow a more extensive investigation of the molecular mechanisms involved in protein-ligand association and dissociation. It would be interesting to see if future work using such an approach will validate our findings here.

The binding of saquinavir to wildtype and mutant HIV-1 proteases has also been investigated, herein, from a thermodynamic perspective using MD techniques. The compromise between applying computationally demanding 'exact' free energy methods and the inaccuracies of more 'approximate' methods has been discussed. We have successfully applied an 'approximate' free energy method in calculating the absolute and relative binding free energies of saquinavir to the wildtype and a range of mutant HIV-1 proteases, obtaining excellent correlation with experimental results. A decomposition of the entropic and enthalpic components of binding have enabled a more detailed understanding of the interplay between enthalpy and entropy in the alteration of binding strengths exhibited across mutants. Our findings confirm that longer than conventionally exploited timescales, up to or in excess of 10 ns, are necessary to ensure the convergence of free energy calculations on ligand-protease systems.

The determination of accurate binding free energies reported here, given the success of our method, serves as the first part of the much larger objective that aims to successfully rank the resistance conferred by a substantial range of drug-protease combinations, using the same method. Considerable time and effort is involved in the process of constructing molecular dynamics simulations. To circumvent this, we have developed a tool, the 'Binding Affinity Calculator' (BAC), for fully automating the workflow involved in a HIV-1 protease-ligand binding free energy calculation (see Appendix A). Utilising HPC and grid technology, the BAC thus has the capacity to routinely determine multiple arrays of such binding free energy calculations.

An attractive goal would be the accurate determination of the binding affinity of a range of inhibitors for the unique viral genotypic constitution of an infected individual. Such information, if available on a clinically relevant timescale (less than two weeks), would assist in the optimisation of treatment on a 'patient-specific' basis. Provided the method reported here is able to sensitively rank an arbitrary selection of drug-protease combinations, the infrastructure afforded by the BAC can easily be utilised for such a purpose, well within a two week timescale.

We have applied the same methodology in determining the binding free energy of natural substrates to HIV-1 protease variants, allowing for insights into the alteration of catalytic efficiency induced by mutations. Our findings suggest that an even longer timescale is likely to be necessary to guarantee convergence of the binding free energy for natural substrates, which are larger and more flexible than inhibitors. Mutations can also either decrease or increase the catalytic efficiency of the protease and a combined treatment of the interplay of changes in drug resistance with changes in catalytic efficiency provides an enhanced description of the overall effect of a mutation on the enzymatic fitness of the protease. Based on work by Gulnik *et al.* [224], we have devised a metric for assessing the approximate enzymatic fitness of a mutant relative to the wildtype, computable from free energy methods based on molecular simulation. Future studies, involving possibly longer simulations allowing the convergence of the substrate binding free energy, coupled with the BAC infrastructure, would allow an enzymatic fitness potential to be generated using MD simulations.

Ultimately, the mutational pathways naturally selected by HIV in response to inhibitor treatment emerge from the complexity of the interactions of the virus in its biological environment [170]. Addressing such spatiotemporal scales is hopelessly beyond the reach of current MD capabilities. Instead, the application of hierarchical models, in which MD fulfils one component of several interconnected models that traverse the required spatiotemporal scales necessary to describe a complex system [271], seem more promising.

There are a number of projects currently underway that make use of multi-scale modelling approaches in the attempt to understand complex biological systems. Funded by the BBSRC, a multi-scale modelling project, IntBioSim, is attempting to establish a computational approach to investigate a range of biological processes spanning from the chemical to the sub-cellular level[1]. It therefore includes a molecular description including the QM/MM methods described briefly in § 2.8.1, classical molecular dynamics as well as the development of course-grained molecular dynamics in which groups of atoms are treated as single entities.

Integrative Biology, an EPSRC funded e-Science pilot project, is attempting to take the multi-scale modelling concept further, by developing hierarchical multi-scale models that describe the behaviour of a system from a very macroscopic level such as the mammalian heart, down to the behaviour of ion channels that are associated with its function[2] [271].

An EU funded 6th Framework Project (FP6), known as ViroLab, is attempting to develop a virtual laboratory that can be used by researchers and medical doctors to enhance the treatment of infectious diseases[3]. Although not conventionally a multi-scale modelling project, one aspect of ViroLab is to incorporate molecular level information using molecular dynamics techniques into a central decision support system for the enhancement of clinical anti-retroviral treatment of patients infected with HIV. Further discussion of such an integration scheme can be found in Appendix A.

Finally, an EU funded 7th Framework Project (FP7), known as the Virtual Physiological Human (VPH) is perhaps the most ambitious prospect for multi-scale modelling yet considered. It is emerging from the efforts of another FP6, the Strategy for the European Physiome (STEP) Road Map project[4], which aims to to provide recommendations to the EU for the approach to be adopted in the FP7 VPH Initiative which is running from 2007-2013. The European Physiome or 'EuroPhysiome' is a consortium of various 'physiome' projects, currently banded together under the STEP initiative, in which various integrated models of components of organisms are being studied at several scales. When developed, the VPH aims to be an integrated framework of numerous models that will facilitate the investigation of the human body as a single complex system.

Some of the projects mentioned here are examples of exciting scientific frameworks which may be

[1] http://www.intbiosim.org
[2] http://www.integrativebiology.ac.uk
[3] http://www.virolab.org
[4] http://www.europhysiome.org/roadmap

used to develop new multi-scaled models that better understand the complex biological and biochemical processes relating to HIV. For example, there are currently no models which describe the overall enzymatic fitness of the protease in terms of the interaction of all of its cleaved substrates in the presence of inhibitors. Nor is there a model of how the combined set of reaction rates of differing viral proteins with respective ligands leads to a measure of the overall fitness of a particular viral strain. Interestingly, studies on the 'swarm' or quasi-species of viral strains that infect an individual [232] have shown that the most catalytically efficient enzyme is not necessarily the most abundant, illustrating the highly non-linear aspect of such systems. The development of an interconnected set of models for determining the viral fitness of HIV in its combined physiological and chemotherapeutic environment would be an interesting and beneficial challenge for the scientific community. Indeed it is one that may be specifically pursued *inter alia* as part of the above-mentioned EU funded ViroLab and Virtual Physiological Human projects. The enzymatic fitness would then be determined by molecular simulation; this in turn would be a factor in determining the fitness of the viral strain, which would subsequently contribute to understanding the fitness of the viral quasi-species in a person.

A great milestone in the treatment of HIV would be the ability to predetermine the mutational pathway traversed by the virus in attaining resistance. This would enable strategies to be developed to 'outwit' the virus by controlling its evolution in a desired direction. The successful fusion of deductive methods such as molecular dynamics, inductive bioinformatical methods based on clinical and empirical observations [272] and integrative multi-scale models are likely to contribute to this. The exponential rate of increase in computational power will facilitate the use of the MD techniques reported here in such a synergistic process.

However, we end this thesis on a more philosophical note. Viruses like HIV, are but a subset of an ever evolving and increasingly complicated biological ecosystem. If the complete understanding of a biological system requires the incorporation of a description of its environment, then it is not at all trivial to know where one should draw a line around the system. Taken to its extreme, one would require a description of all universal interactions. Instead, understanding the immensely complex self-organising nature of biological systems, in general, may require a paradigm shift in our understanding of physics and especially statistical mechanics. Indeed, in the words of Erwin Schrödinger, when referring to the ordered functioning of biological organisms [273], "...we are obviously faced with events whose regular and lawful unfolding is guided by a 'mechanism' entirely different from the 'probability mechanism' of physics." Whether it emerges from advances in non-linear dynamics, fractal geometry or even an alteration of our notions of entropy, the development of such a new physical framework would not only dramatically advance our understanding of complex self-organising systems, but also lead to major applications in the medical domain.

APPENDIX A

The Binding Affinity Calculator (BAC)

Here we describe a tool, called the 'Binding Affinity Calculator' (BAC), developed for the automation of binding affinity calculations of HIV-1 protease-ligand variants. We discuss the motivations for, and the current scope of, this tool as well as the workflow, architecture and methodology adopted by the BAC in the determination of free energies of binding from molecular simulation.

A.1 Design and Scope of the BAC

With the advent of grid technology and the availability of high performance computing (HPC) resources, the opportunity to perform large numbers of CPU intensive simulations has become realistic. Within this context, the design of the BAC is based around the notion that, provided a robust protocol exists for the simulation of a given biomolecular system, a user wanting to study scientific aspects of the system will want to avoid spending time on the repetitive, manual construction and implementation of the required set of MD simulations. Time can be spent more productively if only the varying range of parameters of scientific interest need be specified, whilst an automated protocol exists for the construction of the simulation-ready model and implementation of the resulting set of simulations and post-production analyses.

In Chapter 6, we reported a robust MD methodology for the simulation and subsequent calculation of free energies of binding of HIV-1 protease complexes. The BAC builds on this methodology by automating the various model construction, MD simulation and post-production analysis protocols, whilst requiring the specification of only a few biological input parameters. In its most automated sense, only the identity of the ligand and the sequence of a protease, relative to a designated wildtype, need to be assigned. The user can optionally select from a range of initial crystal structures (PDBs) and assign a protonation state to the catalytic dyad of the protease.

The BAC currently affords the possibility to implement binding free energy calculations for all 9 FDA inhibitors of HIV-1 protease as well as 7 of the natural substrates cleaved by the protease (see Chapters 3, 6 and 7). In addition to this, molecular dynamics simulations of the apo-HIV-1 protease can

be implemented using over 200 potential starting crystal structures.

A.2 Workflow of a Free Energy Calculation

Let us consider the workflow involved for a binding free energy calculation of a HIV-1 protease-ligand variant, which uses the methods described in Chapter 6. We will begin with the assumption that a starting crystal structure of the complex exists and that forcefield and charge parameters for the protein and ligand are also provided. Figure A.1 shows the various steps required for the execution of the workflow.

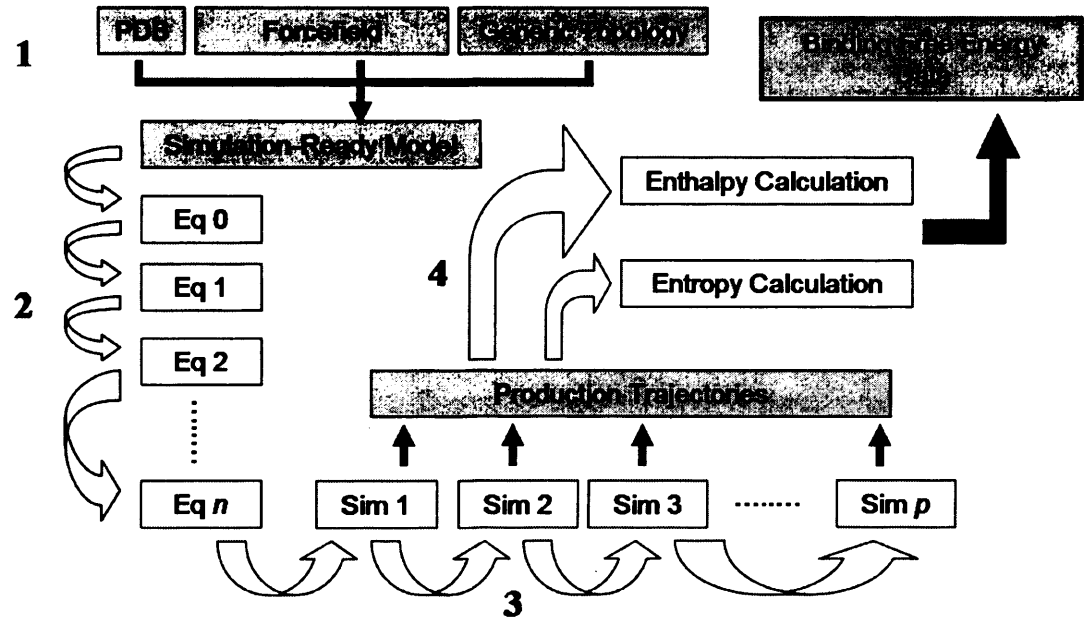


Figure A.1: The workflow of an MMPBSA free energy calculation comprising four sequential stages. **1.** Preparation of a simulation-ready model from the pdb, forcefield parameters and generic topology information. **2.** A linear chain of equilibration simulations. **3.** A linear chain of production simulations each generating trajectories for analysis. **4.** Post-production execution of the enthalpy and entropy calculations leading to a determination of the binding free energy.

Prior to any molecular dynamics, a simulation-ready model has to be generated from the pdb coordinate, generic topology and forcefield parameter information. The process of generating such a model requires the extraction of suitable protease and ligand coordinates, incorporation of any mutations, the addition of neutralising ions and solvation of the target structure. System-specific topology and coordinate files then have to be generated which are ready to be simulated.

The next stage involves the array of sequential equilibration simulations that need to run before production simulations can begin. These include the stages of minimisation, annealing the system,

the gradual relaxing of constraints, which vary based on the mutations that have been incorporated and, finally, unrestrained equilibration in a desired thermodynamic ensemble (see § 6.2). Each step of this sequential protocol utilises a separate configuration file containing the exact instructions for that simulation. The output state data of one step in the protocol is then used as the input state of the following step until the end of the equilibration phase.

The production phase is very similar to equilibration and also consists of a chain of sequentially executed simulations. Each stage of the production phase is executed using a separate configuration file, which again reads in the output-state of the previous stage of the simulation. In principle, it is possible to implement only one stage, where a single simulation is run for a sufficiently long time to traverse the entire production phase. In practice however, the queueing regulations for single continuous computations on many high performance computing (HPC) resources make it more sensible to decompose the production simulation into several sequentially run and individually queued components.

Finally, the trajectory information that is output in the production phase is then post-processed in order to calculate the enthalpies and entropies of binding, as discussed in Chapter 6. Each part of the calculation uses separate configuration files which contain the specific instructions pertaining to the energy calculation method.

It is clear from the above description that, although the workflow involved in a calculation is rather long, the procedures that need to be implemented across a range of protease-ligand variants are very similar. Automation of such a linear workflow when studying an array of varying complexes thus saves considerable time.

A.3 Architecture and Workflow Management of the BAC

Automation of the workflow described in the previous section requires one to overcome two general obstacles. Firstly, all of the preparation files necessary for a simulation-ready model and associated configuration files necessary for the execution of the chain of simulations, as well as for the post-production calculation of the binding free energy, need to be generated automatically. Secondly, the intensive computational requirement of molecular simulations, in general, requires simulations to be implemented on HPC resources (see Chapter 2). Once the set of files required for a simulation has been generated, these files need to be manually transferred to a HPC resource, where simulations need to be submitted using an appropriate job submission script. After the computation has completed, subsequent output data then needs to be marshalled back to an appropriate storage resource for post-processing.

The architecture of the BAC has been designed in a manner which overcomes these two obstacles to automation (see Figure A.2) and which facilitates its use, in general, across HPC and grid resources. Essential to the full automation conferred by the BAC is the utilisation of the Application Hosting Environment (AHE), discussed in Chapter 2, which manages the workflow around various computational

resources. One aspect of the functionality afforded by the AHE is that job submission can be handled through a command line interface on the client-side resource. Perl scripts can then be used to construct workflows to manage the order in which a series of simulations is conducted. Within the BAC, the 'Unit-Executor' is an example of such a Perl script and is responsible for managing the calculation of a single, uniquely defined protease-ligand sequence, termed a 'unit'. The Unit-Executor is executed from the front-end command line interface of the client-side resource.

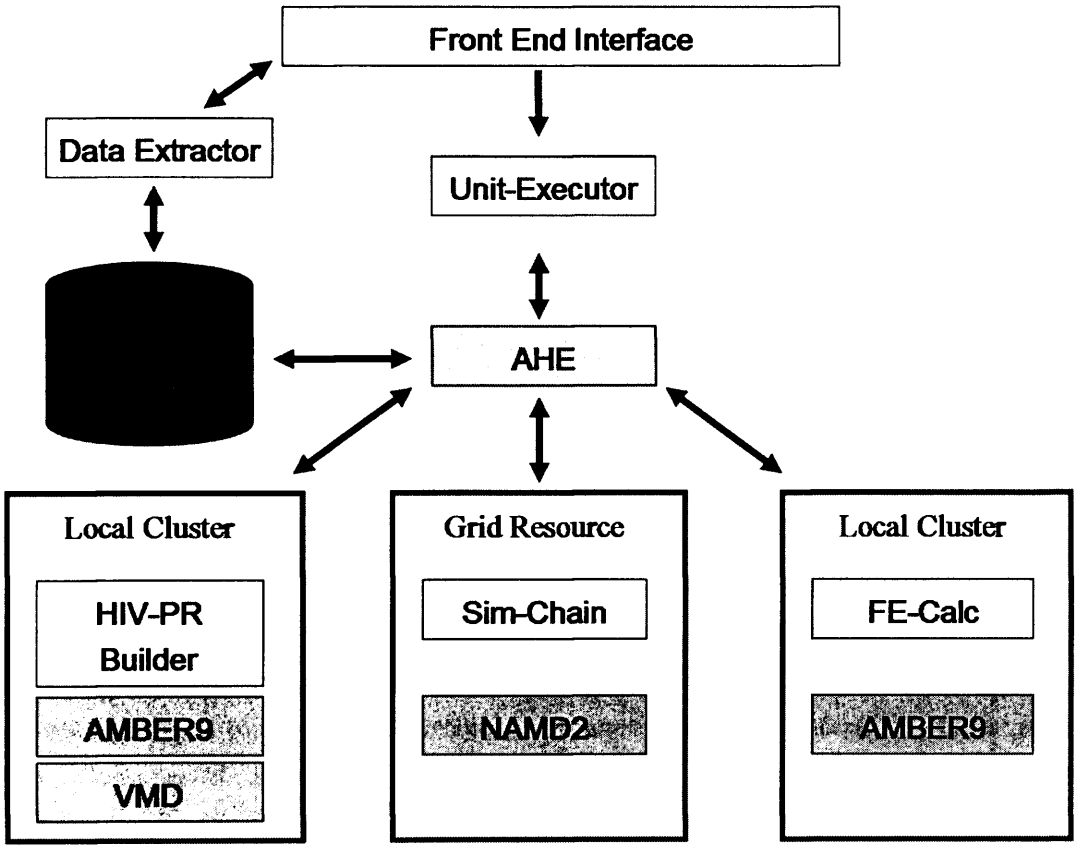


Figure A.2: Architecture of the BAC.

The BAC decomposes the workflow of a complete free energy calculation into three main components, (a) the building of a model, (b) the MD equilibration and simulation of the model and (c) the post-production analysis through which the free energy is calculated. These are implemented by the 'HIV-PR Builder', 'Sim-Chain' and 'FE-Calc' applications respectively. We will describe each of these applications in more detail later, but for now we turn our attention to the management of a single calculation.

Upon initiating the Unit-Executor, the AHE runs the HIV-PR Builder program, typically on a local resource that has the AMBER 9 [53] and VMD [50] software applications installed. The HIV-PR Builder subsequently builds all the pre-simulation files and configuration files necessary for all stages

of the equilibration and simulation, prior to any simulation taking place. In addition, to this it spawns the Sim-Chain program. The AHE then stages all of the required files to a compute resource, including the spawned instance of the Sim-Chain program, which is subsequently executed on that resource. It is necessary for the compute resource to already have the NAMD2 [34] molecular dynamics software, used by Sim-Chain, compiled on it. When each component of the equilibration/simulation run is complete, output data is staged back to a storage resource; the Unit-Executor then checks for successful completion before re-executing the Sim-Chain program for the next component of the simulation. When all stages in the simulation are complete and have been staged back to the data storage resource, the AHE then executes the FE-Calc program, again typically on a local resource. The FE-Calc program generates the input and execution files required for the enthalpy and entropy calculations, implemented respectively using the MMPBSA and normal mode analysis methods described in Chapter 6, and then submits them for calculation. Once the calculations are complete, the calculation output files are staged back to the storage resource and the Unit-Executor terminates. The binding free energy data can then be directly viewed from the storage or extracted in a convenient way using the 'Data Extractor' program, which runs on the front-end command line interface.

The modular design of the BAC allows specific components, such as the HIV-PR Builder, Sim-Chain and FE-Calc applications to be used independently at the cost of complete automation. In such a scenario, the pre-simulation and configuration files required for a specified HIV-1 protease-ligand variant are still automatically generated, affording considerable speed up over manual preparation. However, the user then needs to manually both marshall data from resource to resource and submit jobs. For scientists interested in implementing differences from the default equilibration, simulation and/or free energy calculation protocols, but who wish the relational structure of a set of simulations to be preserved, this can be of great benefit.

A.4 The HIV-PR Builder and Sim-Chain Applications

The HIV-PR Builder application automates the preparation of a simulation-ready molecular dynamics model of HIV protease, either in complex with a ligand or in the apo form. It consists of a set of Perl scripts which include the generation and execution of 'tcl' scripts in the VMD application and 'tleap' commands in the AMBER 9 software package.

To correctly run the HIV-PR Builder, which is executed from the command line, it is necessary to specify the forcefield, the initial pdb crystal structure, the complexed status of the protease (either drug-bound, substrate-bound or apo), the ligand identity, if bound, and the protonation state of the catalytic dyad. In addition, optional parameters with default values may be specified, such as any desired mutations relative to the crystal structure chosen and/or mutations relative to the peptide substrate selected, as well as the size of the solvation box.

The builder contains a host of over 200 pre-modified pdb structures of HIV protease with atomic nomenclature in the AMBER format and the two chains of the protease designated A and B sequentially. Atomic coordinates have been left unaltered. All of these pdbs can be used as the basis for apo-protease structures. There are a restricted set of complexed pdb structures required to generate structures of the 9 corresponding FDA inhibitor and 7 natural substrate complexes available for binding free energy calculations. The coordinates of drugs and substrates have been pre-extracted from these structures. Furthermore, drug charge parameters have also been predetermined using protocols similar to those described in the 'Methods' section of Chapter 4.

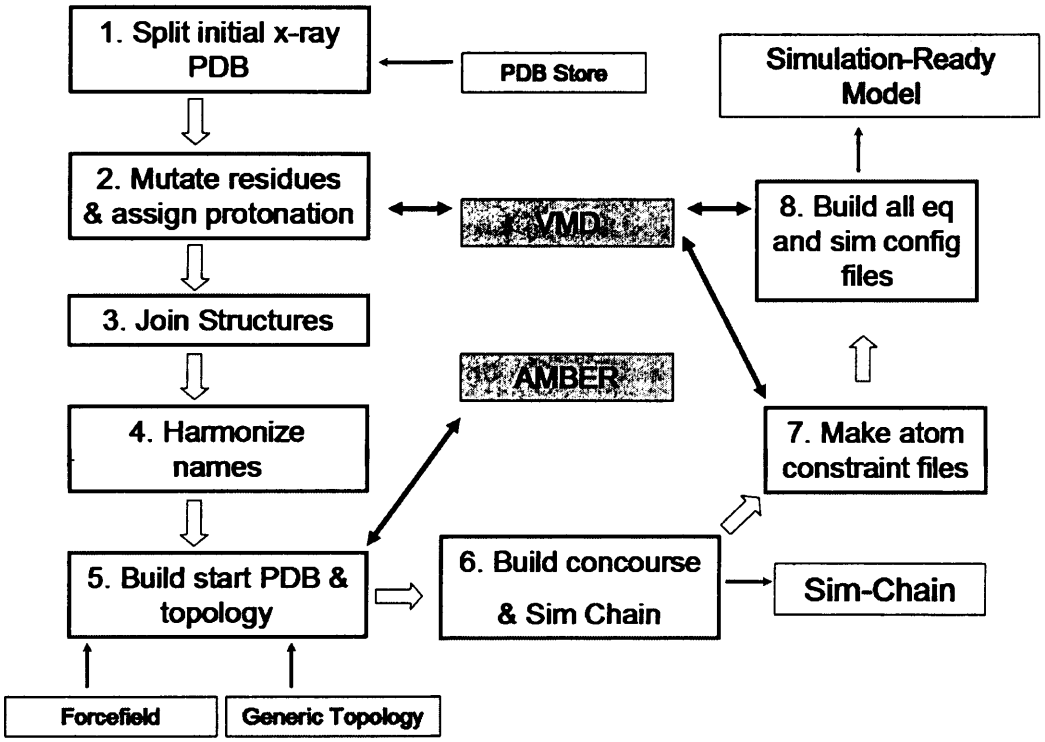


Figure A.3: Schematic representation of the HIV-PR Builder application.

Figure A.3 shows a schematic representation of the processes implemented by the builder. Once executed, the HIV-PR Builder splits the two monomeric chains (A and B) of the protease in the pdb specified into separate coordinate files alongside any ions and all crystallographic water molecules. If a substrate has been specified, this too is extracted into another file. Each of these files are then subject to the incorporation of mutations by means of a 'tcl' script run on the VMD command line interface and generated by a Perl script. The protonation of the dyad is similarly assigned. The separate coordinate files are merged and atomic nomenclature, previously assigned by VMD, is converted back into AMBER nomenclature. A source file, with instructions to add a drug, neutralising ions and water

molecules as well as construct starting topology and coordinate files, is generated by a Perl script and subsequently executed using the 'tleap' module of AMBER 9. Following this, the main directory, termed the 'concourse' and its sub-directories are generated, into which all subsequent data corresponding to the specified unique HIV-1 protease-ligand combination will be stored. Simulation start files are transferred to a concourse sub-directory. The Sim-Chain application resides within the HIV-PR Builder and consists of a collection of modified job submission scripts for a range of HPC resources. It is subsequently copied to another concourse sub-directory.

The minimisation, equilibration and simulation implemented by the BAC makes use of NAMD2. All equilibration and simulation stages require individual configuration files in the NAMD format. Furthermore, several components of the equilibration stage require constraint files to be accessible. These specify the atoms in the system that will be constrained with a certain force constant (see Chapter 2). However, as all of the details of equilibration and simulation configuration are pre-determinable and follow the protocols reported in Chapter 6, the builder generates all constraint files and configuration files at this stage. 'Tcl' scripts are generated and executed by a Perl script using the VMD command line interface to construct the appropriate constraint files.

Generation of equilibration and simulation configuration files proceeds as follows. The cell basis vectors are computed using a 'tcl' script in VMD; these are then used to determine optimal PME values for the initial stage of the equilibration. Temperature, pressure and constraint settings are automatically written to each configuration file as well as the number of simulation timesteps. The number of equilibration stages varies according to the number of mutants incorporated into a system. For each mutation, there is an additional equilibration configuration file during which the constraints around the mutation are relaxed (see Chapter 6). However the total number of timesteps for the whole equilibration phase remains constant (2 ns). Input/output files are specified in the configuration file in a systematic manner which ensures that the output of one stage is named as the input of the following and all file-paths are assigned relative to the concourse directory. The only differences in the simulation configuration files are the names of the input/output files which are written in a similar systematic manner.

Once generation of all pre-simulation files is complete, the modular design of each specific protease-ligand unit, contained entirely within its respective concourse directory, facilitates its transfer to different compute resources. The Sim-Chain application can then be run by executing each individual job submission script, from within the appropriate concourse sub-directory. This is possible as the job submission scripts also utilise a naming scheme relative to the concourse unit. Each submission script is designed to sequentially submit a range of equilibration/simulation stages, executed by NAMD2 and, by default, using 32 processors on the host compute resource. When not called by the AHE, the user must initiate each script manually after checking that the set of simulations has terminated correctly. When interfaced with the AHE, the Unit-Executor uses the AHE to check for successful completion, before execution of the following job submission script.

A.5 The FE-Calc Application

The FE-Calc application executes MMPBSA and normal mode analysis calculations using the MMPBSA module of the AMBER 9 software package and consists of a Perl script that generates all the input files necessary for a calculation, subsequently submitting these calculations to the compute resource.

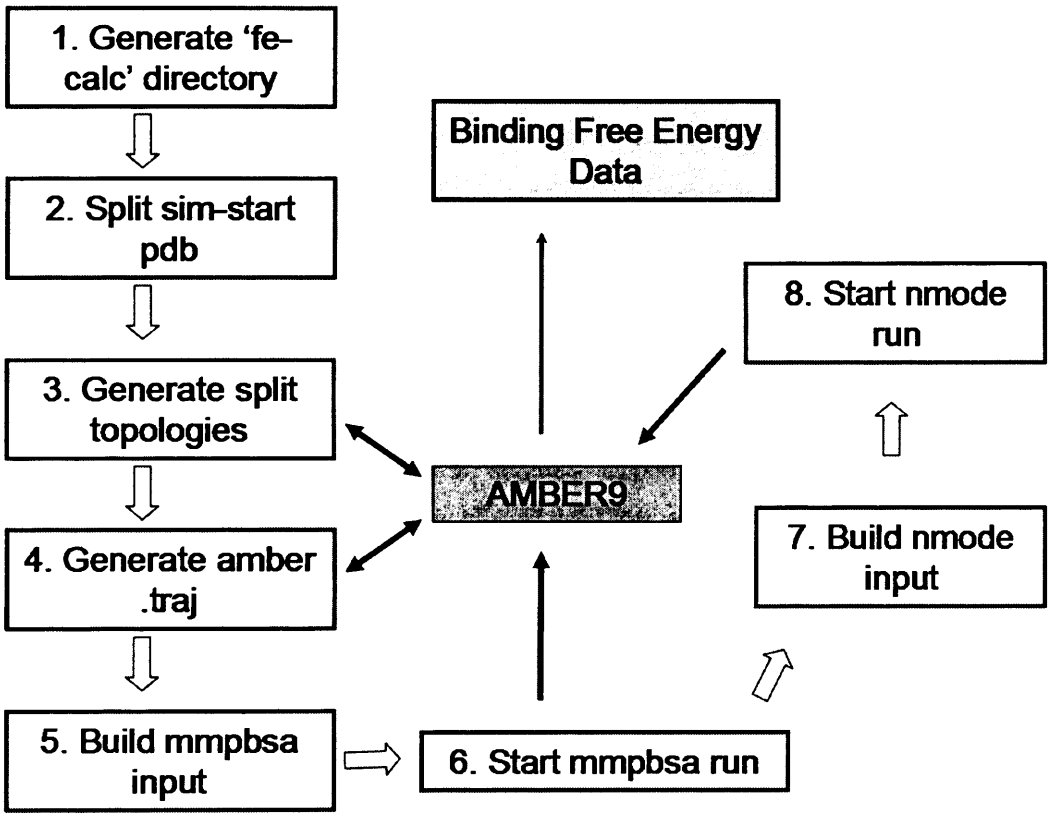


Figure A.4: Schematic representation of the FE-Calc application.

Figure A.4 shows the processes implemented by the application. FE-Calc takes in similar input to the HIV-PR Builder application and uses this to identify the unique concourse unit to process. An 'fe-calc' concourse sub-directory is produced and all subsequently generated files are written therein.

The MMPBSA module requires separate topology files to be written for the complex, ligand and receptor as well as input trajectories written in the AMBER .traj format. The simulation-ready starting pdb for the original molecular dynamics simulation is split into 3 separate pdbs, for the complex, ligand and receptor. These are used to generate separate topologies using a 'tleap' source file written by the Perl script. Amber trajectories are generated by executing source files written by FE-Calc for the PTRAJ module of AMBER 9. MMPBSA and normal mode analyses use different parameters and are thus implemented from separate input files. These are generated from existing templates and subsequently

modified by the application. FE-Calc then determines appropriate atom numbers for the beginning and end of each molecular species and assigns these along with the snapshot frequency and output filenames to each input file. Generic job submission scripts are then used to launch both the MMPBSA and NMODE calculations on the compute resource.

A.6 The Clinical Motivation for a Binding Affinity Calculator

From the perspective of clinical medicine, there are two fundamental problems that generate an imposing barrier in the quest to prescribe the most effective treatment for patients infected with HIV. These stem from the large genetic variability of the retrovirus as well as the emergence of characteristic mutations associated with the treatment of many available anti-retroviral inhibitors (ARVs) [173].

One prevalent clinical problem has been ascertaining which drug regimen best treats a patient's viral genotype. Whilst genotypic assaying of individuals infected with HIV is a standard procedure implemented to obtain portions of the viral sequence [274], the interpretation of such information, given the complexity of emergent mutational patterns [170], both in treatment-naive and treated individuals, often means that clinicians have to resort to 'decision- support' software for assistance [275]. Such 'decision-support' applications use existing clinical databases as well as phenotypic information from the conduction of inhibition studies to infer the susceptibility of a range of inhibitors to a particular viral sequence.

A second problem is that, even though an initially optimal drug regimen may be determined through such a process, as well as some insight provided on the mutational response of inhibitor treatment, predicting the mutational pathway of the viral genotype for any given genotype and for any given inhibitor, remains a challenge. Invariably, the effect of treatment has been the subsequent emergence of resistance mutations that impair inhibitor efficacy, causing a subsequent evolution of the viral genotype within the host and allowing the viral load to increase. Monitoring of treatment efficacy leads to a re-evaluation of the optimal drug regimen based on the evolved viral genotype and this in turn can lead to an alteration in treatment. Unfortunately, evolution of the viral genotype can result in mutational pathways that lead to multi-drug resistant (MDR) viruses [135, 204].

The efficacy of any 'decision-support' software is ultimately dependent on the extent of available clinical or phenotypic information relevant to a particular sequence. Unfortunately, the determination of drug binding affinities either by experimental or computational means, which is by no means trivial, has conventionally taken far too long to be of immediate use in clinical response. Instead studies on binding affinity are constrained to provide information only in retrospect, once drug resistant mutations have evolved in viral populations and have been characterised clinically. Whilst such studies, once conducted, are invaluable for optimising treatment in clinical response to characterised mutations, they are not informative about mutations that have not been characterised, but which may exist in infected

individuals.

The ViroLab project [276, 277], discussed at the end of Chapter 2, is attempting to overcome such problems. In addition to providing a common virtual environment for the collective accessibility of a range of previously independent clinical cohorts, one aim of ViroLab is to incorporate computational data at the molecular level, complementing existing clinical and phenotypic data. Any lack of significant data that may exist for the interpretation of an optimal treatment against a unique viral genotype may then be bridged by conducting a suitable computational study on that particular genotype.

There are several requirements for a tool that can bridge such a gap. Firstly, it should be predictive, thus not requiring previous clinical information as input. Secondly, it should be accurate when ranking the susceptibility of inhibitors to a variety of viral genotypes. Furthermore, it should be automated so that the clinician need not be concerned with the specific methodology of a calculation and, finally, it should return on appropriate timescales (< 2 weeks) to be of use in assisting with a clinical response. Such a ranking tool, provided it can meet these requirements, can in principle be integrated into an expert system such as ViroLab, providing complementary information to existing clinical data.

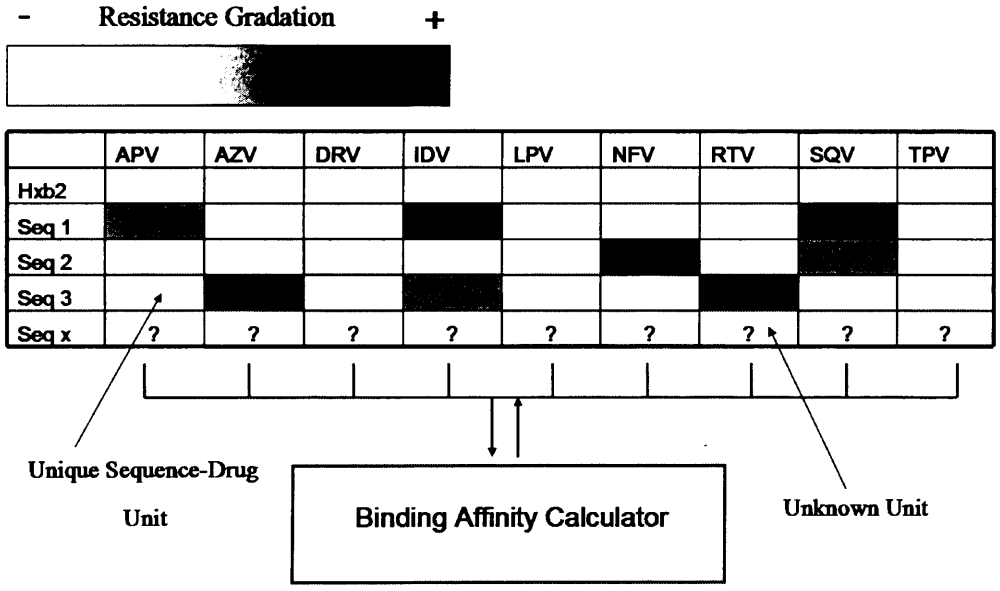


Figure A.5: Illustration of a BAC determinable resistance profile.

The BAC fulfils all of these criteria. Although not so far tested on a clinically large array of protease-ligand variants, the excellent quantitative ranking of drug resistant mutants exhibited in the study conducted in Chapter 6 is encouraging for future studies on different drug-bound protease variants. An illustration of how the BAC could, in the future, be used to grade the resistance conferred by a unique drug-protease sequence is shown in Figure A.5. The determination of the resistance conferred by a

unique protease-drug combination relative to a wildtype structure (e.g HXB2) is naturally determined from a single BAC 'unit'. Although represented similarly to grading schemes used by existing databases [275], the fundamental difference is that grading would be determined quantitatively from free energy calculations, in a routine and automated fashion.

APPENDIX B

Internal Coordinate, Structure and Partial Atomic Charge Information for Saquinavir

In Table B.1 we report the contents of the 'prep'file generated by the ANTECHAMBER module of AMBER, attained for saquinavir using the RESP procedure described in § 4.2.1. This file contains the internal coordinate information, the structural connectivity and the partial atomic charge information for saquinavir. The values listed here were used for all studies reported in Chapters 4-6.

0 0 2

This is a remark line

molecule.res

SAQ XYZ 0

CORRECT OMIT DU BEG

0.0000

1	DUMM	DU	M	0	-1	-2	0.000	.0	.0	.00000
2	DUMM	DU	M	1	0	-1	1.449	.0	.0	.00000
3	DUMM	DU	M	2	1	0	1.522	111.1	.0	.00000
4	C1	c3	M	3	2	1	1.540	111.208	180.000	-0.396
5	H1	hc	E	4	3	2	1.084	97.263	-141.648	0.094
6	H2	hc	E	4	3	2	1.086	149.716	4.607	0.094
7	H3	hc	E	4	3	2	1.087	78.479	111.432	0.094
8	C2	c3	M	4	3	2	1.531	42.004	-29.266	0.570
9	C3	c3	3	8	4	3	1.533	109.656	-17.237	-0.396
10	H4	hc	E	9	8	4	1.086	109.969	-60.490	0.094
11	H5	hc	E	9	8	4	1.086	110.508	58.791	0.094
12	H6	hc	E	9	8	4	1.080	111.052	179.352	0.094
13	C4	c3	3	8	4	3	1.535	109.538	-138.635	-0.396
14	H7	hc	E	13	8	4	1.086	110.262	61.311	0.094

15	H8	hc	E	13	8	4	1.081	110.897	-178.577	0.094
16	H9	hc	E	13	8	4	1.086	110.518	-58.135	0.094
17	N1	n	M	8	4	3	1.472	106.157	102.413	-0.374
18	H10	hn	E	17	8	4	0.995	117.634	3.448	0.187
19	C5	c	M	17	8	4	1.345	126.361	179.163	0.466
20	O1	o	E	19	17	8	1.207	123.951	5.087	-0.545
21	C6	c3	M	19	17	8	1.532	116.355	-170.627	0.133
22	H11	h1	E	21	19	17	1.083	103.659	-170.022	0.031
23	C7	c3	M	21	19	17	1.564	107.142	76.290	-0.139
24	H12	hc	E	23	21	19	1.085	107.751	111.959	0.037
25	H13	hc	E	23	21	19	1.085	110.017	-3.061	0.037
26	C8	c3	M	23	21	19	1.533	115.058	-127.472	0.077
27	H14	hc	E	26	23	21	1.090	106.269	-163.073	0.018
28	C9	c3	M	26	23	21	1.535	115.530	79.417	-0.042
29	H15	hc	E	28	26	23	1.088	109.499	57.566	0.004
30	H16	hc	E	28	26	23	1.085	110.350	-60.338	0.004
31	C10	c3	M	28	26	23	1.532	110.912	179.203	-0.001
32	H17	hc	E	31	28	26	1.089	109.475	-64.071	-0.004
33	H18	hc	E	31	28	26	1.087	110.110	179.013	-0.004
34	C11	c3	M	31	28	26	1.531	110.988	56.788	0.055
35	H19	hc	E	34	31	28	1.087	110.270	-177.763	0.003
36	H20	hc	E	34	31	28	1.088	109.025	65.932	0.003
37	C12	c3	M	34	31	28	1.532	111.359	-55.815	-0.183
38	H21	hc	E	37	34	31	1.089	108.632	-67.256	0.040
39	H22	hc	E	37	34	31	1.087	110.255	176.551	0.040
40	C13	c3	M	37	34	31	1.537	112.936	53.276	0.018
41	H23	hc	E	40	37	34	1.088	107.215	-167.230	0.016
42	C14	c3	M	40	37	34	1.543	111.866	75.439	0.072
43	H24	h1	E	42	40	37	1.084	108.510	96.942	0.064
44	H25	h1	E	42	40	37	1.084	110.273	-18.150	0.064
45	N2	n3	M	42	40	37	1.461	114.480	-137.644	-0.286
46	C15	c3	M	45	42	40	1.458	113.146	172.023	-0.249
47	H26	h1	E	46	45	42	1.088	105.819	-63.381	0.094
48	H27	h1	E	46	45	42	1.082	108.729	-177.050	0.094
49	C16	c3	M	46	45	42	1.543	124.974	56.819	0.199
50	H28	h1	E	49	46	45	1.084	101.705	-175.702	0.068

51	O2	oh	S	49	46	45	1.409	110.109	-60.644	-0.710
52	H29	ho	E	51	49	46	0.948	108.465	-164.111	0.434
53	C17	c3	M	49	46	45	1.542	118.597	69.604	0.579
54	H30	h1	E	53	49	46	1.083	108.030	38.656	0.034
55	C18	c3	3	53	49	46	1.536	114.344	-84.397	-0.297
56	H31	hc	E	55	53	49	1.080	108.867	44.408	0.085
57	H32	hc	E	55	53	49	1.083	108.699	-69.882	0.085
58	C19	ca	S	55	53	49	1.518	114.196	167.146	0.124
59	C20	ca	B	58	55	53	1.394	120.762	-64.268	-0.180
60	H33	ha	E	59	58	55	1.076	120.418	0.622	0.127
61	C21	ca	B	59	58	55	1.383	120.647	-178.908	-0.137
62	H34	ha	E	61	59	58	1.076	119.477	179.970	0.139
63	C22	ca	B	61	59	58	1.388	120.412	0.127	-0.151
64	H35	ha	E	63	61	59	1.075	120.274	-179.972	0.132
65	C23	ca	B	63	61	59	1.382	119.369	0.040	-0.137
66	H36	ha	E	65	63	61	1.076	120.096	-179.991	0.139
67	C24	ca	S	65	63	61	1.387	120.218	-0.086	-0.180
68	H37	ha	E	67	65	63	1.076	119.348	-179.696	0.127
69	N3	n	M	53	49	46	1.467	108.611	155.699	-0.982
70	H38	hn	E	69	53	49	0.999	118.565	121.478	0.431
71	C25	c	M	69	53	49	1.344	122.929	-74.743	0.787
72	O3	o	E	71	69	53	1.208	123.752	8.281	-0.660
73	C26	c3	M	71	69	53	1.537	115.158	-174.232	0.106
74	H39	h1	E	73	71	69	1.082	107.405	-48.218	0.070
75	C27	c3	3	73	71	69	1.531	110.513	-165.818	-0.297
76	H40	hc	E	75	73	71	1.084	108.152	174.642	0.094
77	H41	hc	E	75	73	71	1.082	108.974	57.615	0.094
78	C28	c	B	75	73	71	1.520	114.158	-67.187	0.919
79	O4	o	E	78	75	73	1.203	120.982	-59.289	-0.625
80	N4	n	B	78	75	73	1.357	116.267	125.392	-1.062
81	H42	hn	E	80	78	75	0.998	114.230	-171.373	0.435
82	H43	hn	E	80	78	75	0.996	116.723	-33.890	0.435
83	N5	n	M	73	71	69	1.447	111.965	69.264	-0.378
84	H44	hn	E	83	73	71	1.001	118.590	93.008	0.193
85	C29	c	M	83	73	71	1.336	122.149	-88.671	0.571
86	O5	o	E	85	83	73	1.212	123.664	3.754	-0.619

87	C30	ca	M	85	83	73	1.509	115.382	-176.405	0.289
88	C31	ca	M	87	85	83	1.416	118.875	177.845	-0.272
89	H45	ha	E	88	87	85	1.071	119.348	-0.460	0.178
90	C32	ca	M	88	87	85	1.357	117.981	179.463	-0.159
91	H46	ha	E	90	88	87	1.076	120.804	179.936	0.167
92	C33	ca	M	90	88	87	1.417	119.692	-0.138	0.064
93	C34	ca	M	92	90	88	1.419	123.436	-179.647	-0.240
94	H47	ha	E	93	92	90	1.075	119.032	0.079	0.161
95	C35	ca	M	93	92	90	1.358	120.228	-179.994	-0.146
96	H48	ha	E	95	93	92	1.075	120.097	-179.873	0.149
97	C36	ca	M	95	93	92	1.418	120.522	0.042	-0.091
98	H49	ha	E	97	95	93	1.075	119.264	179.979	0.142
99	C37	ca	M	97	95	93	1.357	120.538	-0.020	-0.332
100	H50	ha	E	99	97	95	1.075	122.060	-179.969	0.200
101	C38	ca	M	99	97	95	1.419	120.016	-0.077	0.445
102	N6	nb	M	101	99	97	1.351	118.577	-179.798	-0.498

LOOP

N2	C6
C13	C8
C24	C19
N6	C30
C38	C33

IMPROPER

C6	N1	C5	O1
C18	C20	C19	C24
C19	C21	C20	H33
C20	C22	C21	H34
C21	C23	C22	H35
C22	C24	C23	H36
C19	C23	C24	H37
C26	N3	C25	O3
C27	N4	C28	O4
C30	N5	C29	O5
C29	C31	C30	N6

C30	C32	C31	H45
C31	C33	C32	H46
C38	C32	C33	C34
C33	C35	C34	H47
C34	C36	C35	H48
C35	C37	C36	H49
C38	C36	C37	H50
C33	C37	C38	N6

DONE

STOP

Table B.1: Partial atomic charges, internal coordinate and structure information for saquinavir as contained in the 'prep' file generated by the RESP procedure in the ANTECHAMBER module of AMBER. These values were used for all studies reported in Chapters 4-6. Coloumns 1-3 describe atom number, name and type respectively. Coloumn 4 describes chain designation, whilst coloumns 5-7 describe structural connectivity. Coloumns 8-10 describe equilibrium bond length, angle and dihedral angle parameters whilst coloumn 11 describes the partial atomic charges.

Bibliography

[1] *The Holy Qur'an.*

[2] Heisenberg, W., 2000. *Physics and Philosophy.* Penguin Classics.

[3] Alberts, B., D. Bray, J. Lewis, M. Raff, K. Roberts and J. D. Watson, 2002. *Molecular Biology of the Cell.* Garland Publishing Inc., Fourth edition.

[4] Stryer, L., J. M. Berg and J. L. Tymoczko, 2002. *Biochemistry.* W. H. Freeman and Co. Ltd., Fifth edition.

[5] Branden, C. and J. Tooze, 1999. *Introduction to Protein Structure.* Garland Publishing Inc., Second edition.

[6] Price, N. C., R. A. Dwek, R. G. Ratcliffe and M. R. Wormald, 2001. *Principles and Problems in Physical Chemistry for Biochemists.* Oxford Univeristy Press, Third edition.

[7] Anfinsen, C. B., 1973. Principles that govern the folding of protein chains. *Science* 181:223–230.

[8] Borgnia, M., S. Nielsen, A. Engel and P. Agre, 1999. Cellular and molecular biology of aquaporin water channels. *Annual Review of Biochemistry* 68:425.

[9] Orengo, C. A., J. M. Thornton and D. T. Jones, 2002. *Bioinformatics: Genes, Proteins and Computers.* Bios Scientific Publishers Ltd.

[10] Levinthal, C., 1968. Are there pathways for protein folding? *Journal of Chemical Physics* 65:44–45.

[11] Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, 2000. The protein data bank. *Nucleic Acids Research* 28:235–242.

[12] Fowler, P. W. and P. V. Coveney, 2006. A computational protocol for the integration of the monotopic protein prostaglandin H2 synthase into a phospholipid bilayer. *Biophysical Journal* 91:401–410.

[13] Fowler, P. W., K. Balali-Mood, S. Deol, P. V. Coveney and M. S. P. Sansom, 2007. Monotopic enzymes and lipid bilayers: A comparative study. *Biochemistry* 46:3108–3115.

[14] Adkins, C. J., 1983. *Equilibrium Thermodynamics*. Cambridge Univeristy Press, Third edition.

[15] Maschera, B., G. Darby, G. Palu, L. L. Wright, M. Tisdale, R. Myers, E. D. Blair and E. S. Furfine, 1996. Human immunodeficiency virus: Mutations in the viral protease that confer resistance to saquinavir increase the dissociation rate constant of the protease-saquinavir complex. *Journal of Biological Chemistry* 271:33 231–33 235.

[16] Chen, X., Y. Lin and M. K. Gilson, 2001. A web-accessible molecular recognition database. *Journal of Combinatorial Chemistry and High-Throughput Screening* 4:719–725.

[17] Chen, X., Y. Lin and M. K. Gilson, 2002. The binding database: Overview and user's guide. *Biopolymers Nucleic Acid Science* 61:127–141.

[18] Tajkhorshid, E., P. Nollert, M. Jensen, L. J. W. Miercke, J. O'Connell, R. M. Stroud and K. Schulten, 2002. Control of the selectivity of the aquaporin water channel family by global orientational tuning. *Science* 296:525–530.

[19] Larios, E., J. S. Li, K. Schulten, H. Kihara and M. Gruebele, 2004. Multiple probes reveal a native-like intermediate during low-temperature refolding of ubiquitin. *Journal of Molecular Biology* 340:115–125.

[20] Allen, M. P. and D. J. Tildesley, 1987. *Computer Simulation of Liquids*. Oxford Science Publications.

[21] Frenkel, D. and B. Smit, 2002. *Understanding Molecular Simulation: From Algorithms to Applications*. Academic Press Inc., Second edition.

[22] Leach, A. R., 2001. *Molecular Modelling: Principles and Applications*. Prentice-Hall, Second edition.

[23] Landau, L. D. and E. M. Lifshitz, 1980. *Mechanics*, volume 1 of *Course of Theoretical Physics*. Butterworth-Heinemann Ltd., Third edition.

[24] Wang, J. M., P. Cieplak and P. A. Kollman, 2000. How well does a restrained electrostatic potential (RESP) model perform in calculating conformational energies of organic and biological molecules? *Journal of Computational Chemistry* 21:1049–1074.

[25] Cornell, W. D., P. Cieplak, C. I. Bayly, I. R. Gould, K. M. Merz, Jr., D. M. Ferguson, D. C. Spellmeyer, T. Fox, J. W. Caldwell and P. A. Kollman, 1995. A second generation force field for the simulation of proteins, nucleic acids and organic molecules. *Journal of the American Chemical Society* 117(19):5179–5197.

[26] Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, 1983. CHARMM: A program for macromolecular energy, minmimization, and dynamics calculations. *Journal of Computational Chemistry* 4:187–217.

[27] Ott, K.-H. and B. Meyer, 1996. Parametrization of GROMOS force field for oligosaccharides and assessment of efficiency of molecular dynamics simulations. *Journal of Computational Chemistry* 17:1068–1084.

[28] Wang, J., R. M. Wolf, D. A. Case and P. A. Kollman, 2004. Development and testing of a general AMBER force field (GAFF). *Journal of Computational Chemistry* 25:1157–1174.

[29] Verlet, L., 1967. Computer 'experiments' on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Physical Review* 159:98–103.

[30] Potter, D., 1972. *Computational Physics*. Wiley, New York.

[31] Swope, W. C., H. C. Anderson, P. H. Berens and K. R. Wilson, 1982. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *Journal of Chemical Physics* 76:637–649.

[32] Pearlman, D. A., D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, III, S. DeBolt, D. Ferguson, G. Seibel and P. Kollman, 1995. AMBER, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules. *Computer Physics Communications* 91:1–41.

[33] Case, D. A., T. Cheatham, T. Darden, H. Gohlke, R. Luo, K. M. Merz, Jr., A. Onufriev, C. Simmerling, B. Wang and R. Woods, 2005. The Amber biomolecular simulation programs. *Journal of Computational Chemistry* 26:1668–1688.

[34] Kale, L., R. Skeel, M. Bhandarkar, R. Brunner, A. Gursoy, N. Krawetz, J. Phillips, A. Shinozaki, K. Varadarajan and K. Schulten, 1999. NAMD2: Greater scalability for parallel molecular dynamics. *Journal of Computational Physics* 151:283–312.

[35] Smith, W. H. and T. R. Forester, 1996. A general-purpose parallel molecular dynamics simulation package. *Journal of Molecular Graphics* 14:136–141.

[36] Plimpton, S. J., 1995. Fast parallel algorithms for short-range molecular dynamics. *Journal of Computational Physics* 117:1–19.

[37] Bouzida, D., S. Kumar and R. H. Swendsen, 1992. Efficient Monte Carlo methods for the computer simulation of biological molecules. *Physical Review A* 45:8894–8901.

[38] Ryckaert, J. P., G. Ciccotti and H. J. C. Berendsen, 1977. Numerical integration of the Cartesian equations of motion of a system with constraints: Molecular dynamics of n-alkanes. *Journal of Computational Physics* 23:327–341.

[39] Verlet, L., 1967. Computer 'experiments' on classical fluids. II. Equilibrium correlation functions. *Physical Review* 165:201–204.

[40] Hein, J., L. Smith, I. Bush, M. Guest and P. Sherwood, 2005. On the performance of molecular dynamics applications on current high-end systems. *Philosophical Transactions of the Royal Society A* 363(1833):1987–1998.

[41] Landau, L. D. and E. M. Lifshitz, 1980. *Statistical Physics: Part 1*, volume 5 of *Course of Theoretical Physics*. Butterworth-Heinemann Ltd., Third edition.

[42] Parthia, R. K., 1996. *Statistical Mechanics*. Butterworth Heinemann, Second edition.

[43] Anderson, H. C., 1980. Molecular dynamics simulations at constant pressure and/or temperature. *Journal of Chemical Physics* 72:2384–2393.

[44] Nosé, S. A., 1984. A unified formulation of the constant temperature molecular-dynamics methods. *Journal of Chemical Physics* 81(1):511–519.

[45] Nosé, S. A., 1984. A molecular-dynamics method for simulations in the canonical ensemble. *Molecular Physics* 52(2):255–268.

[46] Hoover, W. G., 1985. Canonical dynamics - equilibrium phase-space distributions. *Physical Review A* 31(3):1695–1697.

[47] Adelman, S. A. and J. D. Doll, 1979. Generalized Langevin equation approach for atom-solid-surface scattering: General formulation for classical scattering off harmonic solids. *Journal of Chemical Physics* 64(6):2375–2388.

[48] Adelman, S. A., 1979. Generalized Langevin theory for many-body problems in chemical dynamics - general formulation and the equivalent harmonic chain representation. *Journal of Chemical Physics* 71:4471–4486.

[49] Berendsen, H. J. C., J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, 1984. Molecular dynamics with coupling to an external bath. *Journal of Chemical Physics* 81:3684–3690.

[50] Humphrey, W., A. Dalke and K. Schulten, 1996. VMD - Visual molecular dynamics. *Journal of Molecular Graphics* 14:33–38.

[51] Balsera, M. A., W. Wriggers, Y. Oono and K. Schulten, 1996. Principal component analysis and long time protein dynamics. *Journal of Physical Chemistry* 100:2567–2572.

[52] Caves, L. S. D., J. D. Evanseck and M. Karplus, 1998. Locally accessible conformations of proteins: Multiple molecular dynamics simulations of crambin. *Protein Science* 7:649–666.

[53] Case, D. A., T. A. Darden, T. E. Cheatham, III, C. L. Simmerling, J. Wang, R. E. Duke, R. Luo, K. M. Merz, D. A. Pearlman, M. Crowley, R. C. Walker, W. Zhang, B. Wang, S. Hayik, A. Roitberg, G. Seabra, K. F. Wong, F. Paesani, X. Wu, S. Brozell, V. Tsui, H. Gohlke, L. Yang, C. Tan, J. Mongan, V. Hornak, G. Cui, P. Beroza, D. H. Mathews, C. Schafmeister, W. S. Ross and P. A. Kollman, 2006. AMBER 9. University of California, San Francisco.

[54] Mongan, J., 2004. Interactive essential dynamics. *Journal of Computer-Aided Molecular Design* 18:433–436.

[55] Isralewitz, B., M. Gao and K. Schulten, 2001. Steered molecular dynamics and mechanical functions of proteins. *Current Opinion in Structural Biology* 11:224–230.

[56] Israilev, S., S. Stepaniants, M. Balsera, Y. Oono and K. Schulten, 1997. Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophysical Journal* 72:1568–1581.

[57] Isralewitz, B., S. Israilev and K. Schulten, 1997. Binding pathway of retinal to bacterio-opsin: A predicition by molecular dynamics simulations. *Biophysical Journal* 73:2972–2979.

[58] Wriggers, W. and K. Schulten, 1999. Investigating a back door mechanism of actin-phosphate release by steered molecular dynamics. *Proteins: Structure, Function and Genetics* 35:262–273.

[59] Shen, L., J. Shen, X. Luo, F. Cheng, Y. Xu, K. Chen, E. Arnold, J. Ding and H. Jiang, 2003. Steered molecular dynamics simulation on the binding of NNRTI to HIV-1 RT. *Biophysical Journal* 84:3547–3563.

[60] Wang, W., O. Donini, C. M. Reyes and P. A. Kollman, 2001. Biomolecular simulations: Recent developments in force fields, simulations of enzyme catalysis, protein-ligand, protein-protein, and protein-nucleic acid noncovalent interactions. *Annual Review of Biophysics and Biomolecular Structure* 30:211–243.

[61] Fowler, P. W., S. Jha and P. V. Coveney, 2005. Grid-based steered thermodynamic integration accelerates the calculation of binding free energies. *Philosophical Transactions of the Royal Society A* 363:1999–2015.

[62] Aqvist, J., C. Medina and J.-E. Samuelsson, 1994. A new method for predicting binding affinity in computer-aided drug design. *Protein Engineering* 7:385–391.

[63] Hansson, T. and J. Aqvist, 1995. Estimation of binding affinities for HIV proteinase inhibitors by molecular dynamics simulations. *Protein Engineering* 8:1137–1144.

[64] Hansson, T., J. Marelius and J. Aqvist, 1998. Ligand binding affinity prediction by linear interaction energy methods. *Journal of Computer-Aided Molecular Design* 12:27–35.

[65] Wang, W., J. Wang and P. A. Kollman, 1999. What determines the van der Waals coefficient β in the LIE (linear interaction energy) method to estimate binding free energies using molecular dynamics simulations. *Proteins: Structure, Function and Genetics* 34:395–402.

[66] Carlson, H. A. and W. L. Jorgensen, 1995. An extended linear response method for determining free energies of hydration. *Journal of Physical Chemistry* 99:10 667–10 673.

[67] Sanner, M. F., A. J. Olson and J. C. Spehner, 1996. Reduced surface: an efficient way to compute molecular surfaces. *Biopolymers* 38(3):305–320.

[68] Sitkoff, D., K. A. Sharp and B. Honig, 1994. Accurate calculation of hydration free-energies using macroscopic solvent models. *Journal of Physical Chemistry* 98(7):1978–1988.

[69] Tidor, B. and M. Karplus, 1994. The contribution of vibrational entropy to molceular association. *Journal of Molecular Biology* 238:405–414.

[70] Schwarzl, S. M., T. B. Tschopp, J. C. Smith and S. Fischer, 2002. Can the calculation of ligand binding free energies be improved with continuum solvent electrostatics and an ideal-gas entropy correction? *Journal of Computational Chemistry* 23:1143–1149.

[71] Srinivasan, J., T. E. Cheatham, III, P. Cieplak, P. A. Kollman and D. A. Case, 1998. Continuum solvent studies of the stability of DNA, RNA, and phosphoramidate-DNA helices. *Journal of the American Chemical Society* 120(37):9401–9409.

[72] Brooks, B. R., D. Janezic and M. Karplus, 1995. Harmonic analysis of large systems. I. Methodology. *Journal of Computational Chemistry* 16(12):1522–1542.

[73] Andricioaei, I. and M. Karplus, 2001. On the calculation of entropy from covariance matrices of the atomic fluctuations. *Journal of Chemical Physics* 115(14):6289–6292.

[74] Cui, Q. and I. Bahar, 2005. *Normal Mode Analysis: Theory and Applications to Biological and Chemical Systems*. Chapman and Hall.

[75] Blood, P. D. and G. A. Voth, 2006. Direct observation of Bin/amphiphysin/Rvs (BAR) domain-induced membrane curvature by means of molecular dynamics simulations. *Proceedings of the National Academy of Sciences* 103(41):15 068–15 072.

[76] Foster, I. and C. Kesselman, 2004. *The Grid 2: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann, Second edition.

[77] Coveney, P. V., 2005. Scientific grid computing. *Philosophical Transactions of the Royal Society A* 363(1833):1707–1713.

[78] Foster, I. and C. Kesselman, 1997. Globus: A metacomputing infrastructure toolkit. *International Journal of Supercomputer Applications* 11(2):115–128.

[79] Almond, J. and D. Snelling, 1999. UNICORE: Uniform access to supercomputing as an element of electronic commerce. *Future Generation Computer Systems* 613:1–10.

[80] Chin, J. and P. V. Coveney, 2004. Towards tractable toolkits for the grid: A plea for lightweight usable middleware. *UK e-Science Technical Report* URL `http://nesc.ac.uk/technicalpapers/UKeS-2004-01.pdf`.

[81] Kewley, J., R. Allan, R. Crouchley, D. Grose, T. van Ark, M. Haynes and L. Morris, 2005. GROWL: A lightweight grid services toolkit and applications. *Proceedings of the 4th UK e-Science All Hands Meeting* URL `http://www.allhands.org.uk/2005/proceedings/papers/460.pdf`.

[82] Coveney, P. V., J. Vicary, J. Chin and M. Harvey, 2005. WEDS: A WSRF-based environment for distributed simulation. *Philosophical Transactions of the Royal Society A* 363:1807–1816.

[83] Coveney, P. V., R. S. Saksena, S. J. Zasada, M. McKeown and S. Pickles, 2007. The application hosting environment: Lightweight middleware for grid-based computational science. *Computer Physics Communications* 176:406–418.

[84] Harting, J., J. Chin, M. Venturoli and P. V. Coveney, 2005. Large-scale lattice Boltzmann simulations of complex fluids: Advances through the advent of computational grids. *Philosophical Transactions of the Royal Society A* 363:1895–1915.

[85] Jha, S., P. V. Coveney and M. J. Harvey, 2006. SPICE: Simulated pore interactive computing environment - using federated grids for "Grand Challenge" biomolecular simulations. *International Supercomputer Conference* .

[86] Jarzynski, C., 1997. Nonequilibrium equality for free energy differences. *Physical Review Letters* 78(14):2690–2693.

[87] Jarzynski, C., 2002. Targeted free energy perturbation. *Physical Review E* 65(4):046 122.

[88] Marti-Renom, M. A., A. C. Stuart, A. Fiser, R. Sanchez, F. Melo and A. Sali, 2000. Comparative protein structure modeling of genes And genomes. *Annual Review of Biophysics and Biomolecular Structure* 29:291–325.

[89] Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953. Equation of state calculations by fast computing machines. *Journal of Chemical Physics* 21:1087–1092.

[90] Frisch, M. J., G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, V. G. Zakrzewski, J. A. Montgomery, Jr., R. E. Stratmann, J. C. Burant, S. Dapprich, J. M. Millam, A. D. Daniels, K. N. Kudin, M. C. Strain, O. Farkas, J. Tomasi, V. Barone, M. Cossi, R. Cammi, B. Mennucci, C. Pomelli, C. Adamo, S. Clifford, J. Ochterski, G. A. Petersson, P. Y. Ayala, Q. Cui, K. Morokuma, P. Salvador, J. J. Dannenberg, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. Cioslowski, J. V. Ortiz, A. G. Baboul, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. Gomperts, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, J. L. Andres, C. Gonzalez, M. Head-Gordon, E. S. Replogle, and J. A. Pople, 2001. Gaussian 98, Revision A.10. Gaussian, Inc., Pittsburgh, PA, 2001.

[91] Carloni, P., U. Rothlisberger and M. Parrinello, 2002. The role and perspective of ab initio molecular dynamics in the study of biological systems. *Accounts of Chemical Research* 35:455–464.

[92] Lee, Y. S., S. E. Worthington, M. Krauss and B. R. Brooks, 2002. Reaction mechanism of chorismate mutase studied by the combined potentials of quantum mechanics and molecular mechanics. *Journal of Physical Chemistry B* 106(46):12 059–12 065.

[93] Laio, A., J. VandeVondele and U. Rothlisberger, 2002. A Hamiltonian electrostatic coupling scheme for hybrid Car-Parrinello molecular dynamics simulations. *Journal of Chemical Physics* 116:6941–6947.

[94] Csanyi, G., T. Albaret, M. C. Payne and A. D. Vita, 2004. 'Learn on the Fly': A hybrid classical and quantum-mechanical molecular dynamics simulation. *Physical Review Letters* 93(17).

[95] Delgado-Buscalioni, R. and P. V. Coveney, 2004. Hybrid molecular-continuum fluid dynamics. *Philosophical Transactions of the Royal Society London A* 362:1639–1654.

[96] Delgado-Buscalioni, R. and P. V. Coveney, 2003. USHER: An algorithm for particle insertion in dense fluids. *Journal of Chemical Physics* 119:978–987.

[97] Fabritiis, G. D., R. Delgado-Buscalioni and P. V. Coveney, 2004. Energy controlled insertion of polar molecules in dense fluids. *Journal of Chemical Physics* 121:12 139–12 142.

[98] Lynch, G. C. and B. M. Pettitt, 2000. Semi-grand canonical molecular dynamics simulation of bovine pancreatic trypsin inhibitor. *Chemical Physics* 258:405–413.

[99] Masur, H., M. A. Michelis, J. B. Greene, I. Onorato, R. A. Stouwe, R. S. Holzman, G. Wormser, L. Brettman, M. Lange, H. W. Murray and S. Cunningham-Rundles, 1981. An outbreak of community-acquired Pneumocystis carinii pneumonia: Initial manifestation of cellular immune dysfunction. *New England Journal of Medicine* 305:1431–1438.

[100] CDC, 1981. Kaposi's sarcoma and Pneumocystis pneumonia among homosexual men - New York City and California. *Morbidity and Mortality Weekly Report* 30:305–308.

[101] Hunt, R. C., 2007. *Microbiology and Immunology On-line.* University of South Carolina, School of Medicine. URL `http://www.pathmicro.med.sc.edu`.

[102] Stowring, L., A. T. Haase and H. P. Charman, 1979. Serological definition of the lentivirus group of retroviruses. *Journal of Virology* 29:523–528.

[103] Temin, H. M., 1989. Is HIV unique or merely different? *Journal of Acquired Immune Deficiency Syndromes* 2:1–9.

[104] Greene, W. C., 1991. The molecular biology of human immunodeficiency virus type 1 infection. *New England Journal of Medicine* 324:308–17.

[105] Haase, A. T., 1986. Pathogenesis of lentivirus infections. *Nature* 322:130–136.

[106] Franchini, G., C. Gurgo, H.-G. Guo, R. C. Gallo, E. Collalti, K. A. Fargnoli, L. F. Hall, F. Wong-Staal and M. S. Reitz, 1987. Sequence of simian immunodeficiency virus and its relationship to the human immunodeficiency viruses. *Nature* 328:539–543.

[107] Hirsch, V. M., R. A. Olmsted, M. Murphey-Corb, R. H. Purcell and P. R. Johnson, 1989. An African primate lentivirus (SIVsm) closely related to HIV-2. *Nature* 339:389–392.

[108] Alfano, M. and G. Poli, 2004. The HIV life cycle: Multiple targets for antiretroviral agents. *Drug Design Reviews - Online* 1:83–92.

[109] Doms, R. W. and D. Trono, 2000. The plasma membrane as a combat zone in the HIV battlefield. *Genes & Development* 14:2677–2688.

[110] Kaplan, A. H., M. Manchester and R. Swanstrom, 1994. The activity of the protease of human immunodeficiency virus type 1 is initiated at the membrane of infected cells before the release of viral proteins and is required for release To occur with maximum efficiency. *Journal of Virology* 68(10):6782–6786.

[111] Patel, P. H. and B. D. Preston, 1994. Marked infidelity of human immunodeficiency virus type 1 reverse transcriptase at RNA and DNA template ends. *Proceedings of the National Academy of Sciences* 91:549–553.

[112] Ho, D. D., 1997. Dynamics of HIV-1 replication in vivo. *Journal of Clinical Investigation* 99(11):2565–2567.

[113] Wlodawer, A., M. Miller, M. Jaskólski, B. K. Sathyanarayana, E. Baldwin, I. T. Weber, L. M. Selk, L. Clawson, J. Schneider and S. B. H. Kent, 1989. Conserved folding in retroviral proteases: crystal structure of a synthetic HIV-1 protease. *Science* 245:616–621.

[114] Perryman, A. L., J. Lin and J. A. McCammon, 2004. HIV-1 protease molecular dynamics of a wild-type and of the V82F/I84V mutant: Possible contributions to drug resistance and a potential new target site for drugs. *Protein Science* 13:1108–1123.

[115] Jacks, T., M. D. Power, F. R. Massiarz, P. A. Luciw, P. J. Barr and H. E. Varmus, 1988. Charaterization of ribosomal frameshifting in HIV-1 gag-pol expression. *Nature* 331:280–283.

[116] Chatterjee, A., P. Mridula, A. K. Mishra, R. Mittal and R. V. Hosur, 2005. Folding regulates autoprocessing of HIV-1 protease precursor. *Journal of Biological Chemistry* 280(12):11 369–11 378.

[117] Chen, N., A. Morag, N. Almog, I. Blumenzweig, O. Dreazin and M. Kotler, 2001. Extended nucleocapsid protein is cleaved from the gag-pol precursor of human immunodeficiency virus type 1. *Journal of General Virology* 82:581–590.

[118] Pettit, S. C., S. Gulnik, L. E. Everitt and A. H. Kaplan, 2003. The dimer interfaces of protease and extra-protease domains influence the activation of protease and the specificity of gagpol cleavage. *Journal of Virology* 77(1):366–374.

[119] Pettit, S. C., L. E. Everitt, S. Choudhury, B. M. Dunn and A. H. Kaplan, 2004. Initial cleavage of the human immunodeficiency virus type 1 gagpol precursor by its activated protease occurs by an intramolecular mechanism. *Journal of Virology* 78(16):8477–8485.

[120] Pettit, S. C., J. C. Clemente, J. A. Jeung, B. M. Dunn and A. H. Kaplan, 2005. Ordered processing of the human immunodeficiency virus type 1 gagpol precursor is influenced by the context of the embedded viral protease. *Journal of Virology* 79(16):10 601–10 607.

[121] Almog, N., R. Roller, G. Arad, L. Passi-Even, M. A. Wainberg and M. Kotler, 1996. A $p6^{Pol}$-protease fusion protein is present in mature particles of human immunodeficiency virus type 1. *Journal of Virology* 70(10):7228–7232.

[122] Wondrak, E. M. and J. M. Louis, 1996. Influence of flanking sequences on the dimer stability of human immunodeficiency virus type 1 protease. *Biochemistry* 35:12 957–12 962.

[123] Dauber, D. S., R. Ziermann, N. Parkin, D. J. Maly, S. Mahrus, J. L. Harris, J. A. Ellman, C. Petropoulos and C. S. Craik, 2002. Altered substrate specificity of drug-resistant human immunodeficiency virus type 1 protease. *Journal of Virology* 76(3):1359–1368.

[124] Pettit, S. C., N. Sheng, R. Tritch, S. Erickson-Viitanen and R. Swanstrom, 1998. The regulation of sequential processing of HIV-1 gag by the viral protease. *Advances in Experimental Medicine and Biology* 436:15–25.

[125] Pettit, S. C., M. D. Moody, R. S. Wehbie, A. H. Kaplan, P. V. Nantermet, C. A. Klein and R. Swanstrom, 1994. The p2 domain of human immunodeficiency virus type 1 gag regulates sequential proteolytic processing and is required to produce fully infectious virions. *Journal of Virology* 68(12):8017–8027.

[126] Rose, J. R., R. Salto and C. S. Craik, 1993. Regulation of autoproteolysis of the HIV-1 and HIV-2 proteases with engineered amino acid substitutions. *Journal of Biological Chemistry* 268(16):11 939–11 945.

[127] Mildner, A. M., D. J. Rothrock, J. W. Leone, C. A. Bannow, J. M. Lull, I. M. Reardon, J. L. Sarcich, W. J. Howe, C.-S. C. Tomich, C. W. Smith, R. L. Heinrikson and A. G. Tomasseli, 1994. The HIV-1 protease as enzyme and substrate: Mutagenesis of autolysis sites and generation of a stable mutant with retained kinetic properties. *Biochemistry* 33:9405–9413.

[128] Prabu-Jeyabalan, M., E. Nalivaika and C. A. Schiffer, 2000. How does a symmetric dimer recognize an asymmetric substrate? A substrate complex of HIV-1 protease. *Journal of Molecular Biology* 301:1207–1220.

[129] Prabu-Jeyabalan, M., E. Nalivaika and C. A. Schiffer, 2002. Substrate shape determines specificity of recognition for HIV-1 protease: Analysis of crystal structures of six substrate complexes. *Structure* 10:369–381.

[130] Prabu-Jeyabalan, M., E. Nalivaika, N. M. King and C. A. Schiffer, 2003. Viability of a drug-resistant human immunodeficiency virus type 1 protease variant: Structural insights for better antiviral therapy. *Journal of Virology* 10(2):1306–1315.

[131] Prabu-Jeyabalan, M., E. Nalivaika, N. M. King and C. A. Schiffer, 2004. Structural basis for coevolution of a human immunodeficiency virus type 1 nucleocapsid-p1 cleavage site with a V82A drug-resistant mutation in viral protease. *Journal of Virology* 78(22):12 446–12 454.

[132] Zoete, V., O. Michielin and M. Karplus, 2002. Relation between sequence and structure of HIV-1 protease inhibitor complexes: A model system for the analysis of protein flexibility. *Journal of Molecular Biology* 315:21–52.

[133] Piana, S., P. Carloni and M. Parrinello, 2002. Role of conformational fluctuations in the enzymatic reaction of HIV-1 protease. *Journal of Molecular Biology* 319:567–583.

[134] Prabu-Jeyabalan, M., E. Nalivaika, K. Romano and C. A. Schiffer, 2006. Mechanism of substrate recognition by drug-resistant human immunodeficiency virus type 1 protease variants revealed by a novel structural intermediate. *Journal of Virology* 80(7):3607–3616.

[135] Layten, M., V. Hornak and C. Simmerling, 2006. The open structure of a multi-drug-resistant HIV-1 protease is stabilized by crystal packing contacts. *Journal of the American Chemical Society* 128:13 360–13 361.

[136] Scott, W. R. P. and C. A. Schiffer, 2000. Curling of flap tips in HIV-1 protease as a mechanism for substrate entry and tolerance of drug resistance. *Structure* 8:1259–1265.

[137] Kumar, M. and M. V. Hosur, 2003. Adaptability and flexibility of HIV-1 protease. *European Journal of Biochemistry* 270:1231–1239.

[138] Wang, W. and P. A. Kollman, 2000. Free energy calculations on dimer stability of the HIV protease using molecular dynamics and a continuum solvent model. *Journal of Molecular Biology* 303:567–582.

[139] Piana, S., P. Carloni and U. Rothlisberger, 2002. Drug resistance in HIV-1 protease: Flexibility-assisted mechanism of compensatory mutations. *Protein Science* 11:2393–2402.

[140] Kurt, N., W. R. P. Scott, C. A. Schiffer and T. Haliloglu, 2003. Cooperative fluctuations of unliganded and substrate-bound HIV-1 protease: A structure-based analysis on a variety of conformations from crystallography and molecular dynamics simulations. *Proteins: Structure, Function and Genetics* 51:409–422.

[141] Chang, C.-E., T. Shen, J. Trylska, V. Tozzini and J. A. McCammon, 2006. Gated binding of ligands to HIV-1 protease: Brownian dynamics simulations in a coarse-grained model. *Biophysical Journal* 90:3880–3885.

[142] Hornak, V., A. Okur, R. C. Rizzo and C. Simmerling, 2006. HIV-1 protease flaps spontaneously open and reclose in molecular dynamics simulations. *Proceedings of the National Academy of Sciences* 103:915–920.

[143] Ishima, R., D. I. Freedberg, Y. X. Wang, J. M. Louis and D. A. Torchia, 1999. Flap opening and dimer-interface flexibility in the free and inhibitor-bound HIV protease, and their implications for function. *Structure* 7:1047–1055.

[144] Freedberg, D. I., R. Ishima, J. Jacob, Y-X.Wang, I. Kustanovich, J. M. Louis and D. A. Torchia, 2002. Rapid structural fluctuations of the free HIV protease flaps in solution: Relationship to crystal structures and comparison with predictions of dynamics calculations. *Protein Science* 11:221–232.

[145] Rick, S. W., J. W. Erickson and S. K. Burt, 1998. Reaction path and free energy calculations of the transition between alternate conformations of HIV-1 protease. *Proteins: Structure, Function and Genetics* 32:7–16.

[146] Brik, A. and C. Wang, 2003. HIV-1 protease: Mechanism and drug discovery. *Organic & Biomolecular Chemistry* 1:5–14.

[147] Trylska, J., P. Grochowski and J. A. McCammon, 2004. The role of hydrogen bonding in the enzymatic reaction catalyzed by HIV-1 protease. *Protein Science* 13:513–528.

[148] Chatfield, D. C. and B. R. Brooks, 1995. HIV-1 protease cleavage mechanism elucidated with molecular dynamics simulation. *Journal of the American Chemical Society* 117:5561–5572.

[149] Trylska, J., P. Bala, M. Geller and P. Grochowski, 2002. Molecular dynamics simulations of the first steps of the reaction catalyzed by HIV-1 protease. *Biophysical Journal* 83:794–807.

[150] Piana, S. and P. Carloni, 2000. Conformational flexibility of the catalytic Asp dyad in HIV-1 protease: an ab initio study on the free enzyme. *Proteins: Structure, Function and Genetics* 39:26–36.

[151] Piana, S., D. Sebastiani, P. Carloni and M. Parrinello, 2001. Ab initio molecular dynamics-based assignment of the protonation state of pepstatin A/HIV-1 protease cleavage site. *Journal of the American Chemical Society* 123:8730–8737.

[152] Northrop, D., 2001. Follow the protons: A low-barrier hydrogen bond unifies the mechanism of the aspartic proteases. *Accounts of Chemical Research* 34:790–797.

[153] Okimoto, N., M. Hata, T. Hoshino and M. Tsuda, 2000. Protein hydrolysis mechanism of HIV-1 protease: Investigation by the ab initio MO calculations. *Riken Review* 29:100–102.

[154] Park, H., J. Suh and S. Lee, 2000. Ab initio studies on the catalytic mechanism of aspartic proteinases: Nucleophilic versus general acid/general base mechanism. *Journal of the American Chemical Society* 122:3901–3908.

[155] Hyland, L. J., T. A. Tomaszek, Jr. and T. D. Meek, 1991. Human immunodeficiency virus-1 protease. 2. Use of pH rate studies and solvent kinetic isotope effects to elucidate details of chemical mechanism. *Biochemistry* 30:8454–8463.

[156] Kovalskyy, D., V. Dubyna, A. E. Mark and A. Korenelyuk, 2005. A molecular dynamics study of the structural stability of HIV-1 protease under physiological conditions: The role of Na+ ions in stabilizing the active site. *Proteins: Structure, Function and Bioinformatics* 58:450–458.

[157] Chen, X. and A. Tropsha, 1995. Relative binding free energies of peptide inhibitors of HIV-1 protease: The influence of the active site protonation state. *Journal of Medicinal Chemistry* 38:42–46.

[158] Harte, W. E., Jr. and D. L. Beveridge, 1993. Prediction of the protonation state of the active site aspartyl residues in HIV- 1 protease-inhibitor complexes via molecular dynamics simulation. *Journal of the American Chemical Society* 115:3883–3886.

[159] Ido, E., H.-P. Han, F. J. Kezdy and J. Tang, 1991. Kinetic studies of human immunodeficiency virus type 1 protease and its active-site hydrogen bond mutant A28S. *Journal of Biological Chemistry* 266(36):24 359–24 366.

[160] Smith, R., I. M. Brereton, C. R. Y and S. Kent, 1996. Ionization states of the catalytic residues in HIV-1 protease. *Nature Structural Biology* 3:946–950.

[161] Wang, Y.-X., D. I. Freedberg, T. Yamazaki, P. T. Wingfield, S. J. Stahl, J. D. Kaufman, Y. Kiso and D. A. Torchia, 1996. Solution NMR evidence that the HIV-1 protease catalytic aspartyl groups have different ionization states in the complex formed with the Asymmetric drug KNI-272. *Biochemistry* 35(31):9945–9950.

[162] Trylska, J., J. Antosiewicz, M. Geller, C. N. Hodge, R. M. Klabe, M. S. Head and M. K. Gilson, 1999. Thermodynamic linkage between the binding of protons and inhibitors to HIV-1 protease. *Protein Science* 8:180–195.

[163] Yamazaki, T., L. K. Nicholson, D. A. Torchia, P. Wingfield, S. J. Stahl, J. D. Kaufman, C. J. Eyermann, C. N. Hodge, P. Y. S. Lam, Y. Ru, P. K. Jadhav, C.-H. Chang and P. C. Weber, 1994. NMR and X-ray evidence that the HIV protease catalytic aspartyl groups are protonated in the complex formed by the protease and a non-peptide cyclic urea-based inhibitor. *Journal of the American Chemical Society* 116:10 791–10 792.

[164] Xie, D., S. Gulnik, L. Collins, E. Gustchina, L. Suvorov and J. W. Erickson, 1997. Dissection of the pH dependence of inhibitor binding energetics for an aspartic protease: Direct measurement of the protonation states of the catalytic aspartic acid residues. *Biochemistry* 36:16 166–16 172.

[165] Luo, R., M. S. Head, J. Moult and M. K. Gilson, 1998. pKa shifts in small molecules and HIV protease: Electrostatics and conformation. *Journal of the American Chemical Society* 120:6138–6146.

[166] Wlodawer, A. and J. Vondrasek, 1998. Inhibitors of HIV-1 protease: A major success of structure-assissted drug design. *Annual Review of Biophysics and Biomolecular Structure* 27:249–284.

[167] Lam, P. Y., P. K. Jadhav, C. J. Eyermann, C. N. Hodge, Y. Ru, L. T. Bachelor, J. L. Meek, M. J. Otto, M. M. Rayner, Y. N. Wong, C.-H. Chang, P. C. Weber, D. A. Jackson, T. R. Sharpe and S. Erickson-Viitanen, 1994. Rational design of potent, bioavailable, nonpeptide cyclic ureas as HIV protease inhibitors. *Science* 263:380–384.

[168] Andersson, H. O., K. Fridborg, S. Lowgren, M. Alterman, A. Muhlman, M. Bjorsne, N. Garg, I. Kvarnstrom, W. Schaal, B. Classon, A. Karlen, U. H. Danielsson, G. Ahlsen, U. Nillroth, L. Vrang, B. Oberg, B. Samuelsson, A. Hallberg and T. Unge, 2003. Optimization of P1-P3 groups in symmetric and asymmetric HIV-1 protease inhibitors. *European Journal of Biochemistry* 270:1746–1758.

[169] Fontenot, G., K. J. J. C. Cohen, W. R. Gallaher, J. Robinson and R. B. Luftig, 1992. PCR amplification of HIV-1 proteinase sequences directly from lab isolates allows determination of five conserved domains. *Virology* 190:1–10.

[170] Wu, T. D., C. A. Schiffer, M. Gonzales, J. Taylor, R. Kantor, S. Chou, D. Israelski, A. R. Zolopa, W. J. Fessel and R. W. Shafer, 2003. Mutation patterns and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. *Journal of Virology* 77:4836–4847.

[171] Levine, A. J., J. Momand and C. A. Finlay, 1991. The p53 tumour suppressor gene. *Nature* 351:453–456.

[172] Schinazi, R. F., B. A. Larder and J. W. Mellors, 1997. Mutations in retroviral genes associated with drug resistance. *Antiviral News* 5:129–142.

[173] Johnson, V. A., F. Brun-Vezinet, B. Clotet, B. Conway, D. R. Kuritzkes, D. Pillay, J. Schapiro, A. Telenti and D. Richman, 2005. Update of the drug resistance mutations in HIV-1: 2005. *International AIDS Society - USA* 13:51–57.

[174] Hoffman, N. G., C. A. Schiffer and R. Swanstrom, 2003. Covariation of amino acid positions in HIV-1 protease. *Virology* 314:536–548.

[175] Lech, W. J., G. Wang, Y. L. Yang, Y. Chee, K. Dorman, D. McCrae, L. C. Lazzeroni, J. W. Erickson, J. S. Sinsheimer and A. H. Kaplan, 1996. In vivo sequence diversity of the protease of human immunodeficiency virus type 1: Presence of protease inhibitor-resistant variants in untreated subjects. *Journal of Virology* 70(3):2038–2043.

[176] Svicher, V., F. Ceccherini-Silberstein, F. Erba, M. Santoro, C. Gori, M. C. Bellochi, S. Giannella, M. P. Trotta, A. d'Arminio Monforte, A. Antinori and C. F. Perno, 2005. Novel human immunodeficiency virus type 1 protease mutations potentially involved in resistance to protease inhibitors. *Antimicrobial Agents and Chemotherapy* 49:2015–2025.

[177] Yahi, N., C. Tamalet, C. Tourres, N. Tivoli, F. Ariasi, F. Volot, J.-A. Gastaut, H. Gallais, J. Moreau and J. Fantini, 1999. Mutation patterns of the reverse transcriptase and protease genes in human immunodeficiency virus type 1-infected patients undergoing combination therapy: Survey of 787 sequences. *Journal of Clinical Microbiology* 37:4099–4106.

[178] Martinez-Picado, J., M. P. DePasquale, N. Kartonis, G. Hanna, J. Wong, D. Finzi, E. Rosenberg, H. F. Gunthard, L. Sutton, A. Savara, C. J. Petropoulos, N. Hellmann, B. D. Walker, D. D. Richman, R. Silidano and R. T. D'Aquila, 2000. Antiretroviral resistance during successful therapy of HIV type 1 infection. *Proceedings of the National Academy of Sciences* 97(20):10 948–10 953.

[179] Molla, A., M. Korneyeva, Q. Gao, S. Vasavanonda, P. J. Schipper, H.-M. Mo, M. Markowitz, T. Chernyavskiy, P. Niu, N. Lyons, A. Hsu, R. Granneman, D. D. Ho, C. A. B. Boucher, J. M. Leonard, D. W. Norbeck and D. J. Kempf, 1996. Ordered accumulation of mutations in HIV protease confers resistance to ritonavir. *Nature Medicine* 2(7):760–766.

[180] Condra, J. H., D. J. Holder, W. A. Schleif, O. M. Blahy, R. M. Danovich, L. J. Gabryelski, D. J. Graham, D. Laird, J. C. Quintero, A. Rhodes, H. L. Robbins, E. Roth, M. Shivaprakash, T. Yang, J. A. Chodakewitz, P. J. Deutsch, R. Y. Leavitt, F. E. Massari, J. W. Mellors, K. E. Squires, R. T. Steigbigel, H. Tepller and E. A. Emini, 1996. Genetic correlates of in vivo viral resistance to indinavir, a human immunodeficiency virus type 1 protease inhibitor. *Journal of Virology* 70(12):8270–8276.

[181] Brown, A. J. L., B. T. Korber and J. H. Condra, 1999. Associations between amino acids in the evolution of HIV type 1 protease sequences under indinavir therapy. *AIDS Research and Human Retroviruses* 15(3):247–253.

[182] Patick, A. K., M. Duran, Y. Cao, D. Shugarts, M. R. Keller, E. Mazabel, M. Knowles, S. Chapman, D. R. Kuritzkes and M. Markowitz, 1998. Genotypic and phenotypic characterization of human immunodeficiency virus type 1 variants isolated from patients treated with the protease inhibitor nelfinavir. *Antimicrobial Agents and Chemotherapy* 42(10):2637–2644.

[183] Boden, D. and M. Markowitz, 1998. Resistance to human immunodeficiency virus type 1 protease inhibitors. *Antimicrobial Agents and Chemotherapy* 42(11):2775–2783.

[184] Doyon, L., S. Tremblay, L. Borgon, E. Wardrop and M. G. Cordingley, 2005. Selection and characterization of HIV-1 showing reduced susceptibility to the non-peptidic protease inhibitor tipranavir. *Antiviral Research* 68:27–35.

[185] Baxter, J. D., J. M. Schapiro, C. A. B. Boucher, V. M. Kohlbrenner, D. B. Hall, J. R. Scherer and D. L. Mayers, 2006. Genotypic changes in human immunodeficiency virus type 1 protease associated with reduced susceptibility and virologic response to the protease inhibitor tipranavir. *Journal of Virology* 80(21):10 794–10 801.

[186] Meyer, S. D., H. Azijn, D. Surleraux, D. Jochmans, A. Tahri, R. Pauwels, P. Wigerinck and M.-P. de Bethune, 2005. TMC114, a novel human immunodeficiency virus type 1 protease inhibitor active against protease inhibitor-resistant viruses, including a rroad range of clinical isolates. *Antimicrobial Agents and Chemotherapy* 49(6):2314–2321.

[187] Surleraux, D. L. N. G., A. Tahri, W. G. Verschueren, G. M. E. Pille, H. A. de Kock, T. H. M. Jonckers, A. Peeters, S. D. Meyer, H. Azijn, R. Pauwels, M.-P. de Bethune, N. M. King, M. Prabu-Jeyabalan and P. B. T. P. Wigerinck, 2005. Discovery and selection of TMC114, a next generation HIV-1 protease inhibitor. *Journal of Medicinal Chemistry* 48:1813–1822.

[188] Nijhuis, M., C. A. B. Boucher, P. Schipper, T. Leitner, R. Schuurman and J. Albert, 1998. Stochastic processes strongly influence HIV-1 evolution during suboptimal protease-inhibitor therapy. *Proceedings of the National Academy of Sciences USA* 95:14 441–14 446.

[189] Winters, M. A., J. M. Schapiro, J. Lawrence and T. C. Merigan, 1998. Human immunodeficiency virus type 1 protease genotypes and in vitro protease inhibitor susceptibilities of isolates from individuals who were switched to other protease inhibitors after long-term saquinavir treatment. *Journal of Virology* 72(6):5303–5306.

[190] Logsdon, B. C., J. F. Vickrey, P. Martin, G. Proteasa, J. I. Koepke, S. R. Terlecky, Z. Wawrzak, M. A. Winters, T. C. Merigan and L. C. Kovari, 2004. Crystal structures of a multidrug-resistant human immunodeficiency virus type 1 protease reveal an expanded active-site cavity. *Journal of Virology* 78(6):3123–3132.

[191] Johnson, V. A., F. Brun-Vezinet, B. Clotet, B. Conway, D. R. Kuritzkes, D. Pillay, J. Schapiro and D. Richman, 2006. Update of the drug resistance mutations in HIV-1: Fall 2006. *International AIDS Society - USA* 14:125–130.

[192] Tisdale, M., R. E. Myers, B. Maschera, N. R. Parry, N. M. Oliver and E. D. Blair, 1995. Cross-resistance analysis of human immunodeficiency virus type 1 variants individually selected for resistance to five different protease inhibitors. *Antimicrobial Agents and Chemotherapy* 39(8):1704–1710.

[193] Lin, Y., X. Lin, L. Hong, S. Foundling, R. L. Heinrikson, S. Thaisrivongs, W. Leelamanit, D. Raterman, M. Shah, B. M. Dunn and J. Tang, 1995. Effect of point mutations on the kinetics and the inhibition of human immunodeficiency virus type 1 protease: Relationship to drug resistance. *Biochemistry* 34:1143–1152.

[194] Ermolieff, J., X. Lin and J. Tang, 1997. Kinetic properties of saquinavir-resistant mutants of human immunodeficiency virus type 1 protease and their implications in drug resistance in Vivo. *Biochemistry* 36:12 364–12 370.

[195] Rose, R. E., Y.-F. Gong, J. A. Greytok, C. M. Bechtold, B. J. Terry, B. S. Robinson, M. Alam, R. J. Colonno and P.-F. Lin, 1996. Human immunodeficiency virus type 1 viral background plays a major role in development of resistance to protease inhibitors. *Proceedings of the National Academy of Sciences USA* 93:1648–1653.

[196] Patick, A. K., H. Mo, M. Markowitz, K. Appelt, B. Wu, L. Musick, V. Kalish, S. Kaldor, S. Reich, D. Ho and S. Werber, 1996. Antiviral and resistance studies of AG1343, an orally bioavailable inhibitor of human immunodeficiency virus protease. *Antimicrobial Agents and Chemotherapy* 40(2):292–297.

[197] Velazquez-Campoy, A., I. Luque, M. J. Todd, M. Milutinovich, Y. Kiso and E. Freire, 2000. Thermodynamic dissection of the binding energetics of KNI-272, a potent HIV-1 protease inhibitor. *Protein Science* 9:1801–1809.

[198] Velazquez-Campoy, A., Y. Kiso and E. Freire, 2001. The binding energetics of first- and second-generation HIV-1 protease inhibitors: Implications for drug design. *Archives of Biochemistry and Biophysics* 390(2):169–175.

[199] Velazquez-Campoy, A., M. J. Todd, S. Vega and E. Freire, 2001. Catalytic efficiency and vitality of HIV-1 proteases from African viral subtypes. *Proceedings of the National Academy of Sciences* 98(11):6062–6067.

[200] Leavitt, S. and E. Freire, 2001. Direct measurement of protein binding energetics by isothermal titration calorimetry. *Current Opinion in Structural Biology* 11:560–566.

[201] Ohtaka, H., A. Velazquez-Campoy, D. Xie and E. Freire, 2002. Overcoming drug resistance in HIV-1 chemotherapy: The binding thermodynamics of amprenavir and TMC-126 to wildtype and drug resistant mutants of the HIV-1 protease. *Protein Science* 11:1908–1916.

[202] Todd, M. J., I. Luque, A. Velazquez-Campoy and E. Freire, 2000. Thermodynamic basis of resistance to HIV-1 protease inhibition: Calorimetric analysis of the V82F/I84V active site resistant mutant. *Biochemistry* 39:11 876–11 883.

[203] Muzammil, S., P. Ross and E. Freire, 2003. A major role for a set of non-active site mutations in the development of HIV-1 protease drug resistance. *Biochemistry* 42:631–638.

[204] Ohtaka, H., A. Schon and E. Freire, 2003. Multidrug resistance to HIV-1 protease inhibition requires cooperative coupling between distal mutations. *Biochemistry* 42:13 659–13 666.

[205] Clemente, J. C., R. Hemrajani, L. E. Blum, M. M. Goodenow and B. M. Dunn, 2003. Secondary mutations M36I and A71V in the human immunodeficiency virus type 1 protease can provide an advantage for the emergence of the primary mutation D30N. *Biochemistry* 42:15 029–15 035.

[206] Clemente, J. C., R. E. Moose, R. Hemrajani, L. R. S. Whitford, L. Govindasamy, R. Reutzel, R. McKenna, M. Agbandje-McKenna, M. M. Goodenow and B. M. Dunn, 2004. Comparing the accumulation of active- and nonactive-site mutations in the HIV-1 protease. *Biochemistry* 43:12 141–12 151.

[207] Resch, W., N. Parkin, T. Watkins, J. Harris and R. Swanstrom, 2005. Evolution of human immunodeficiency virus type 1 protease genotypes and phenotypes in vivo under selective pressure of the protease inhibitor ritonavir. *Journal of Virology* 79(16):10 638–10 649.

[208] Kagan, R. M., M. D. Shenderovich, P. N. R. Heseltine and K. Ramnarayan, 2005. Structural analysis of an HIV-1 protease I47A mutant resistant to the protease inhibitor lopinavir. *Protein Science* 14:1870–1878.

[209] Xie, D., S. Gulnik, E. Gustchina, B. Yu, W. Shao, W. Qoronfleh, A. Nathan and J. Erickson, 2000. Drug resistance mutations can effect dimer stability of HIV-1 protease at neutral pH. *Protein Science* 8:1702–1707.

[210] Ohtaka, H., S. Muzammil, A. Schon, A. Velazquez-Campoy, S. Vega and E. Freire, 2004. Thermodynamic rules for the design of high affinity HIV-1 protease inhibitors with adaptability to mutations and high selectivity towards unwanted targets. *The International Journal of Biochemistry and Cell Biology* 36:1787–1799.

[211] Vega, S., L. Kang, A. Velazquez-Campoy, Y. Kiso, L. M. Amzel and E. Freire, 2004. A structural and thermodynamic escape mechanism from a drug resistant mutation of the HIV-1 protease. *Proteins: Structure, Function and Bioinformatics* 55:594–602.

[212] Prabu-Jeyabalan, M., N. M. King, E. Nalivaika, G. Heilek-Snyder, N. Cammack and C. A. Schiffer, 2006. Substrate envelope and drug resistance: Crystal structure of RO1 in complex with wild-type human immunodeficiency virus type 1 protease. *Antimicrobial Agents and Chemotherapy* 50(4):1518–1521.

[213] King, N. M., L. Melnick, M. Prabu-Jeyabalan, E. A. Nalivaika, S.-S. Yang, Y. Gao, X. Nie, C. Zepp, D. L. Heefner and C. A. Schiffer, 2002. Lack of synergy for inhibitors targeting a multi-drug-resistant HIV-1 protease. *Protein Science* 11:418–429.

[214] King, N. M., M. Prabu-Jeyabalan, E. A. Nalivaika and C. A. Schiffer, 2004. Combating susceptibility to drug resistance: Lessons from HIV-1 protease. *Chemistry and Biology* 11:1333–1338.

[215] King, N. M., M. Prabu-Jeyabalan, E. A. Nalivaika, P. Wigerinck, M.-P. de Bethune and C. A. Schiffer, 2004. Structural and thermodynamic basis for the binding of TMC114, a next-generation human immunodeficiency virus type 1 protease inhibitor. *Journal of Virology* 78(21):12012–12021.

[216] Rick, S. W., J. W. Erickson and S. K. Burt, 1998. Reaction path and free energy calculations of the transition between alternate conformations of HIV-1 protease. *Proteins: Structure, Function and Genetics* 32:7–16.

[217] Collins, J. R., S. K. Burt and J. W. Erickson, 1995. Flap opening in HIV-1 protease simulated by 'activated' molecular dynamics. *Nature Structural Biology* 2:334–338.

[218] Wang, W. and P. A. Kollman, 2001. Computational study of protein specificity: The molecular basis of HIV-1 protease drug resistance. *Proceedings of the National Academy of Sciences* 98:14937–14942.

[219] Levy, Y., A. Caflisch, J. N. Onuchic and P. G. Wolynes, 2004. The folding and dimerization of HIV-1 protease: Evidence for a stable monomer from simulations. *Journal of Molecular Biology* 340:67–79.

[220] Levy, Y. and A. Caflisch, 2003. Flexibility of monomeric and dimeric HIV-1 protease. *Journal of Physical Chemistry* 107:3068–3079.

[221] Mammano, F., C. Petit and F. Clavel, 1998. Resistance-associated loss of viral fitness in human immunodeficiency virus type 1: phenotypic analysis of protease and gag coevolution in protease inhibitor-treated patients. *Journal of Virology* 72(9):7632–7637.

[222] Deveraux, H. L., V. C. Emery, M. A. Johnson and C. Loveday, 2001. Replicative fitness in vivo of HIV-1 variants with multiple drug resistance-associated mutations. *Journal of Medical Virology* 65:218–224.

[223] Gonzalez, L. M. G., R. M. Brindeiro, R. S. Aguiar, H. S. Pereira, C. M. Abreu, M. A. Soares and A. Tanuri, 2004. Impact of nelfinavir resistance mutations on in vitro phenotype, fitness, and replication capacity of human immunodeficiency virus type 1 with subtype B and C proteases. *Antimicrobial Agents and Chemotherapy* 48(9):3552–3555.

[224] Gulnik, S. V., L. I. Suvorov, B. Liu, B. Yu, B. Anderson, H. Mitsuya and J. W. Erickson, 1995. Kinetic characterization and cross-resistance patterns of HIV- 1 protease mutants selected under drug pressure. *Biochemistry* 34:9282–9287.

[225] Nijhuis, M., R. Schuurman, D. de Jong, J. Erickson, E. Gustchina, J. Albert, P. Schipper, S. Gulnik and C. A. B. Boucher, 1999. Increased fitness of drug resistant HIV-1 protease as a result of acquisition of compensatory mutations during suboptimal therapy. *AIDS* 13:2349–2359.

[226] Schock, H. B., V. M. Garsky and L. C. Kuo, 1996. Mutational anatomy of an HIV-1 protease variant conferring cross-resistance to protease inhibitors in clinical trials. *Journal of Biological Chemistry* 271(50):31 957–31 963.

[227] Martinez-Picado, J., A. V. Savara, L. Sutton and R. T. D'Aquila, 1999. Replicative fitness of protease inhibitor-resistant mutants of human immunodeficiency virus type 1. *Journal of Virology* 73(5):3744–3752.

[228] Mammano, F., V. Trouplin, V. Zennou and F. Clavel, 2000. Retracing the evolutionary pathways of human immunodeficiency virus type 1 resistance to protease inhibitors: Virus fitness in the absence and in the presence of drug. *Journal of Virology* 74(18):8524–8531.

[229] Pettit, S. C., G. J. Henderson, C. A. Schiffer and R. Swanstrom, 2002. Replacement of the P1 amino acid of human immunodeficiency virus type 1 gag processing sites can inhibit or enhance the rate of cleavage by the viral protease. *Journal of Virology* 76(20):10 226–10 233.

[230] Feher, A., I. T. Weber, P. Bagossi, P. Boross, B. Mahalingham, J. M. Louis, T. D. Copeland, I. Y. Torshin, R. W. Harrison and J. Tozser, 2002. Effect of sequence polymorphism and drug resistance on two HIV-1 gag processing sites. *European Journal of Biochemistry* 269:4114–4120.

[231] Rosin, C. D., R. K. Belew, G. M. Morris, A. J. Olson and D. S. Goodsell, 1999. Coevolutionary analysis of resistance-evading peptidomimetic inhibitors of HIV-1 protease. *Proceedings of the National Academy of Sciences USA* 96:1369–1374.

[232] Fernandez, G., B. Clotet and M. A. Martinez, 2007. Fitness landscape of HIV-1 protease quasispecies. *Journal of Virology* 81(5):2485–2496.

[233] Bally, F., R. Martinez, S. Peters, P. Sudre and A. Telenti, 2000. Polymorphism of HIV type 1 gag p7/p1 and p1/p6 cleavage sites: Clinical significance and implications for resistance to protease inhibitors. *AIDS Research and Human Retroviruses* 16(13):1209–1213.

[234] Côté, H. C., Z. L. Brumme and P. R. Harrigan, 2001. Human immunodeficiency virus type 1 protease cleavage site mutations associated with protease inhibitor cross-resistance selected by indinavir, ritonavir, and/or saquinavir. *Journal of Virology* 75(2):589–594.

[235] Zennou, V., F. Mammano, S. Paulous, D. Mathez and F. Clavel, 1998. Loss of viral fitness associated with multiple gag and gag-pol processing defects in human immunodeficiency virus type 1 variants selected for resistance to protease inhibitors in vivo. *Journal of Virology* 72(4):3300–3306.

[236] Zhang, Y.-M., H. Imamichi, T. Imamichi, H. C. Lane, J. Falloon, M. B. Vasudevachari and N. P. Salzman, 1997. Drug resistance during indinavir therapy is caused by mutations in the protease gene and in its gag substrate cleavage sites. *Journal of Virology* 71(9):6662–6670.

[237] Doyon, L., G. Croteau, D. Thibeault, F. Poulin, L. Pilote and D. Lamarre, 1996. Second locus involved in human immunodeficiency virus type 1 resistance to protease inhibitors. *Journal of Virology* 70(6):3763–3769.

[238] Maguire, M. F., R. Guinea, P. Griffin, S. Macmanus, R. C. Elston, J. Wolfram, N. Richards, M. H. Hanlon, D. J. T. Porter, T. Wrin, N. Parkin, M. Tisdale, E. Furfine, C. Petropoulos, B. W. Snowden and J.-P. Kleim, 2002. Changes in human immunodeficiency virus type 1 gag at positions L449 and P453 are linked to I50V protease mutants in vivo and cause reduction of sensitivity to amprenavir and improved viral fitness in vitro. *Journal of Virology* 76(15):7398–7406.

[239] Gatanaga, H., Y. Suzuki, H. Tsang, K. Yoshimura, M. F. Kavlick, K. Nagashima, R. J. Gorelick, S. Mardy, C. Tang, M. F. Summers and H. Mitsuya, 2002. Amino acid substitutions in gag protein at non-cleavage sites are indispensable for the development of a high multitude of HIV-1 resistance against protease inhibitors. *Journal of Biological Chemistry* 277(8):5952–5961.

[240] Myint, L., M. Matsuda, Z. Matsuda, Y. Yokomaku, T. Chiba, A. Okano, K. Yamada and W. Sugiura, 2004. Gag non-cleavage site mutations contribute to full recovery of viral fitness in protease inhibitor-resistant human immunodeficiency virus type 1. *Anitmicrobial Agents and Chemotherapy* 48(2):444–452.

[241] Wlodawer, A. and J. W. Erickson, 1993. Structure-based inhibitors of HIV-1 Protease. *Annual Review of Biochemistry* 62:543–585.

[242] Wittayanarakul, K., O. Aruksakunwong, S. Saen-oon, W. Chantratita, V. Parasuk, P. Sompornpisut and S. Hannongbua, 2005. Insights into saquinavir resistance in the G48V HIV-1 protease: Quantum calculations and molecular dynamic simulations. *Biophysical Journal* 88:867–879.

[243] Schuettelkopf, A. W. and D. M. F. van Aalten, 2004. PRODRG - a tool for high-throughput crystallography of protein-ligand complexes. *Acta Crystallographica* D60:1355–1363.

[244] Case, D. A., D. A. Pearlman, J. W. Caldwell, T. E. Cheatham, III, J. Wang, W. S. Ross, C. Simmerling, T. Darden, K. M. Merz, R. V. Stanton, A. Cheng, J. J. Vincent, M. Crowley, V. Tsui,

H. Gohlke, R. Radmer, Y. Duan, J. Pitera, I. Massova, G. L. Seibel, U. C. Singh, P. Weiner and P. A. Kollman, 2002. AMBER 7. University of California, San Francisco.

[245] Lepsik, M., Z. Kriz and Z. Havlas, 2004. Efficiency of a second-generation HIV-1 protease inhibitor studied by molecular dynamics and absolute binding free energy calculations. *Proteins: Structure, Function and Bioinformatics* 57:279–293.

[246] Garcia, A. E. and K. Y. Sanbonmatsu, 2002. α-Helical stabilization by side chain shielding of backbone hydrogen bonds. *Proceedings of the National Academy of Sciences* 99:2782–2787.

[247] Hornak, V., R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, 2006. Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 65:712–725.

[248] Schafmeister, C. E. A. F., W. S. Ross and V. Romanovski, 1995. LEaP. University of California, San Francisco, CA.

[249] Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, 1983. Comparison of simple potential functions for simulating liquid water. *Journal of Chemical Physics* 79:926–935.

[250] Essmann, U., L. Perera, M. L. Berkowitz and T. Darden, 1995. A smooth particle mesh Ewald method. *Journal of Chemical Physics* 103:8577–9593.

[251] Meagher, K. L. and H. A. Carlson, 2005. Solvation Influences Flap Collapse in HIV-1 Protease. *Proteins: Structure, Function and Bioinformatics* 58:119–125.

[252] Toth, G. and A. Borics, 2006. Flap opening mechanism of HIV-1 protease. *Journal of Molecular Graphics and Modelling* 24:465–474.

[253] Wittayanarakul, K., O. Aruksakunwong, P. Sompornpisut, V. Sanghiran-Lee, V. Parasuk, S. Pinitglang and S. Hannongbua, 2005. Structure, dynamics and solvation of HIV-1 protease/saquinavir complex in aqueous solution and their contributions to drug resistance: Molecular dynamic simulations. *Journal of Chemical Information and Modeling* 45:300–308.

[254] Zoete, V., O. Michielin and M. Karplus, 2003. Protein-ligand binding free energy estimation using molecular mechanics and continuum electrostatics. Application to HIV-1 protease inhibitors. *Journal of Computer-Aided Molecular Design* 17:861–880.

[255] Furfine, E. S., E. D'Souza, K. J. Ingold, J. J. Leban, T. Spector and D. J. T. Porter, 1992. Two-step binding mechanism for HIV protease inhibitors. *Biochemistry* 31:7886–7891.

[256] Toth, G. and A. Borics, 2006. Closing of the flaps of HIV-1 protease induced by substrate binding: A model of a flap closing mechanism in retroviral aspartic proteases. *Biochemistry* 45:6606–6614.

[257] Rose, R. B., C. S. Craik and R. M. Stroud, 1998. Domain flexibility in retroviral proteases: Structural implications for drug resistance mutations. *Biochemistry* 37:2607–2621.

[258] Chen, X., I. T. Weber and R. W. Harrison, 2004. Molecular dynamics simulations of 14 HIV protease mutants in complexes with indinavir. *Journal of Molecular Modeling* 10:373–381.

[259] Bartels, C., A. Widmer and C. Ehrhardt, 2005. Absolute free energies of binding of peptide analogs to the HIV-1 protease from molecular dynamics simulations. *Journal of Computational Chemistry* 26(12):1294–1305.

[260] Kalra, P., T. Reddy and B. Jayaram, 2001. Free energy component analysis for drug design: a case study of HIV-1 protease-inhibitor binding. *Journal of Medicinal Chemistry* 44(25):4325–4338.

[261] Rick, S. W., I. A. Topol, J. W. Erickson and S. K. Burt, 1998. Molecular mechanisms of resistance: Free energy calculations of mutation effects on inhibitor binding to HIV-1 protease. *Protein Science* 7:1750–1756.

[262] Wan, S., P. V. Coveney and D. R. Flower, 2005. Peptide recognition by the T cell receptor: Comparison of binding free energies from thermodynamic integration, Poisson-Boltzmann and linear interaction energy approximations. *Philosophical Transactions of the Royal Society A* 363:2037–2053.

[263] Rizzo, R. C., S. Toba and I. D. Kuntz, 2004. A molecular basis for the selectivity of thiadiazole urea inhibitors with stromelysin-1 and gelatinase-A from generalized Born molecular dynamics simulations. *Journal of Medicinal Chemistry* 47(12):3065–3074.

[264] Kuhn, B. and P. A. Kollman, 2000. Binding of a diverse set of ligands to avidin and streptavidin: an accurate quantitative prediction of their relative affinities by a combination of molecular mechanics and continuum solvent models. *Journal of Medicinal Chemistry* 43(20):786–3791.

[265] Huo, S., J. Wang, P. Cieplak, P. A. Kollman and I. D. Kuntz, 2002. Molecular dynamics and free energy analyses of cathepsin D-inhibitor interactions: insight into structure-based ligand design. *Journal of Medicinal Chemistry* 45(7):1412–1419.

[266] Wang, J., P. Morin, W. Wang and P. A. Kollman, 2001. Use of MM-PBSA in reproducing the binding free energies to HIV-1 RT of TIBO derivatives and predicting the binding mode to HIV-1 RT of efavirenz by docking and MM-PBSA. *Journal of the American Chemical Society* 22(123):5221–5230.

[267] Chang, C. A., W. Chen and M. K. Gilson, 2007. Ligand configurational entropy and protein binding. *Proceedings of the National Academy of Sciences* 104(5):1534–1539.

[268] Weis, A., K. Katebzadeh, P. Soderhjelm, I. Nilsson and U. Ryde, 2006. Ligand affinities predicted with the MM/PBSA method: dependence on the simulation method and the force field. *Journal of Medicinal Chemistry* 49(22):6596–6606.

[269] Turner, D., J. M. Schapiro, B. G. Brenner and M. A. Wainberg, 2004. The influence of protease inhibitor resistance profiles on selection of HIV therapy in treatment-naive patients. *Antiviral Therapy* 9:301–314.

[270] Chen, C., Y. Xiao and L. Zhang, 2005. A directed essential dynamics simulation of peptide folding. *Biophysical Journal* 88:3276–3285.

[271] Coveney, P. V. and P. W. Fowler, 2005. Modelling biological complexity: a physical scientist's perspective. *The Journal of the Royal Society Interface* 2:267–280.

[272] Deforche, K., T. Silander, R. Camacho, Z. Grossman, M. A. Soares, K. V. Laetham, R. Kantor, Y. Moreau, A.-M. Vandamme and on behalf of the non B Workgroup, 2006. Analysis of HIV-1 pol sequences using Bayesian networks: Implications for drug resistance. *Bioinformatics* 22(24):2975–2979.

[273] Schrödinger, E., 1967. *What is Life?* Cambridge University Press.

[274] Snoek, J., C. Riva, K. Steegen, Y. Schrooten, B. Maes, L. Vergne, K. V. Laethem, M. Peeters and A. M. Vandamme, 2005. Optimization of a genotypic assay applicable to all human immunodeficiency virus type 1 protease and reverse transcriptase subtypes. *Journal of Virological Methods* 128:47–53.

[275] Kantor, R., R. Machekano, M. J. Gonzales, K. Dupnik, J. M. Schapiro and R. W. Shafer, 2001. Human immunodeficiency virus reverse transcriptase and protease sequence database: an expanded data model integrating natural language text and sequence analysis programs. *Nucleic Acids Research* 29(1):296–299.

[276] Sloot, P. M. A., A. V. Boukhanovsky, W. Keulen, A. Tirado-Ramos and C. A. Boucher, 2005. A grid-based HIV expert system. *Journal of Clinical Monitoring and Computing* 19:263–278.

[277] Sloot, P. M. A., A. Tirado-Ramos, I. Altintas, M. Bubak and C. Boucher, 2006. From molecule to man: Decision support in individualized E-health. *Computer* 39(11):40–46.